



University of Nairobi
Faculty of Science and Technology
Department of Physics

**Machine Learning Approaches to Cancer Diagnostics in
Humans Based on Laser Raman Microspectrometry of
Human Body Fluids**

By

John Irungu Githaiga

BEd Science (Kenyatta University), MSc Physics (Kenyatta University)

Registration No. I80/84761/2012

A thesis Submitted in the Fulfillment of the Requirements for the Award of
the Degree of Doctor of Philosophy in Physics of the University of Nairobi

January, 2022

Declaration

I declare that this thesis is my original work and has not been submitted elsewhere for examination, award of a degree or publication. Where other people's work or my own work has been used, this has properly been acknowledged and referenced in accordance with the University of Nairobi's requirements.

John Irungu Githaiga (I80 / 84761 / 2012)


Department of Physics

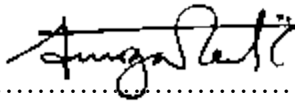
Faculty of Science and Technology

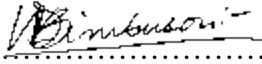
University of Nairobi

Signature *John Githaiga*..... Date ... 20 / 12 / 2021.....

This thesis is submitted for examination with our approval as research supervisors:

| | Signature | Date |
|--|---|----------------|
| Prof. Hudson Kalambuka Angeyo Department of Physics Faculty of Science and Technology University of Nairobi hkalambuka@uonbi.ac.ke |  | 20 / 12 / 2021 |

| | | |
|---|--|-------------------|
| Prof. Kenneth Amiga Kaduki Department of Physics Faculty of Science and Technology University of Nairobi kaduki@uonbi.ac.ke |  | December 21, 2021 |
|---|--|-------------------|

| | | |
|---|---|-------------------|
| Prof. Wallace Dimbuson Bulimo Department of Biochemistry Faculty of Science and Technology University of Nairobi wallace.bulimo@uonbi.ac.ke |  | December 21, 2021 |
|---|---|-------------------|

Dedication

This thesis is dedicated to my family and friends. I thank and admire them for their continued support, encouragement, inspiration and unconditional love.

Acknowledgements

This project was funded in part by The National Council for Science and Technology of Republic of Kenya and the University of Nairobi. It involved a collaboration between the Kenyatta National Hospital (KNH) and Kenya Medical Research Institute (KEMRI). The University of Nairobi funding was provided by the Deans' Committee Research Grant of School of Physical Sciences. The research equipment for spectroscopic analysis was donated by the Swedish International Development Cooperation Agency (SIDA) through the International Science Programme, Uppsala University (ISP). Their contribution is gratefully acknowledged.

I would like to thank my supervisors Dr Hudson Kalambuka Angeyo, Prof. Kenneth Amiga Kaduki and Prof. Wallace Dimbuson Bulimo for their consistent support, patience and critical contribution to this research. Their knowledge has been invaluable in assisting my academic development and understanding.

Prof Wallace Dimbuson Bulimo has been an inspiration. His boundless vitality has rubbed off onto others making a difference in creating imperative collaborations within the medical field. Without his enthusiasm in biophysics, it would have been virtually impossible to achieve significant progress in this project.

Special thanks goes to Dr. Ojuka Daniel Kinyuru (Department of Surgery, University of Nairobi) and Prof. Jessie Nyokabi Githanga (Department of Human Pathology, University of Nairobi) for their time, help and guidance in all pathology related matters. Their acceptance to get involved in this project was the first step towards achieving my objectives. Thanks to Dr. Nyagol Joshua Akelo (Department of Human Pathology, University of Nairobi) and Mr. Justus Okonda (Department of Physics, University of Nairobi) for invaluable guidance in formulating Kenyatta National Hospital – University of Nairobi (KNH – UoN) Ethics and Research Committee proposal.

Dr. Esther Njoki Mwangi Maina of Department of Biochemistry (University of Nairobi) deserves thanks for donating the cell lines. Together with Samwel Lifumo Symekher (Center for Virus Research-KEMRI), they provided professional guidance in preparation of cell lines. They kept

me on track throughout my investigations in spite of occasional long gaps between communications. They have an excellent understanding of cell culture and have taught me to believe in myself as an independent researcher.

During this research I was able to spend many months at the Center for Virus Research-KEMRI. I would like to thank Samwel Lifumo Symekher, Janet Majanja, Meshak Wadegu, Rachila, Shukran, and Silvanos Opanda for their constant kindness and assistance in this project, and for being so welcoming and generous in allowing me to use their resources and intelligence.

Many thanks to Julia Wangui (USAMRD-A) for providing biosafety level 3 training that included bio-risk assessment and identification, specific biosafety measures, code of practice, and general standard operating procedures in biohazards laboratory. The training was invaluable during handling of infectious and hazardous biological materials throughout this project.

At the heart of this research were Peninah Kabethi and Janet Majanja, especially with regard to the extremely important process of collecting and keeping track of biofluid samples and consent forms. Things would have been much more difficult without their medical and organizational abilities.

I am grateful to Mr. Dickson Linen Omucheni and Dr. Zephaniah Birech both at Department of Physics - University of Nairobi for their Raman spectroscopy expertise, without which I would still be in the lab trying to figure out the instrumentation.

For their patience and devotion, I would like to thank my family for helping me through the long days of writing while working full time. I deeply appreciate their love and tolerance over the years. Without their support and encouragement, I would be lost.

Above all, I would like to thank our All-Powerful God. His love and guidance has always provided me with power, good health and courage to tackle all adversities.

Abstract

Near-infrared Raman spectroscopy is a spectroscopic technique capable of providing fingerprint-type information on biochemical molecules. For the early detection of cancer, specific biomarkers, e.g., biofluids' biomarkers, need to be detected with high sensitivity. This enhances diagnostic accuracy in detecting biochemical fingerprints that would point to onset of cancer development. The aim of this study was to test and evaluate novelized machine learning techniques for detection and identification of trace biomarker alterations in saliva and blood pointing to the onset and progression of leukemia and breast cancers via a laser Raman spectral analysis approach. The spectral measurements were done in 393-2063 cm^{-1} region, based on a 785 nm excitation laser. The spectral data analysis were done in the 500-1800 cm^{-1} region; the considered fingerprint region for biological specimens.

Trace biomarkers were studied by analysis of intermediate and higher-order principal components. The utility of intermediate and higher-order principal components in revealing trace biochemical alterations (trace biomarkers) in biological samples was first experimented on prostatic cells' spectra data. The statistical relevance of principal components were determined by the use of the two-sample *t*-test and the effect size statistical criteria. For breast cancer and leukemia studies, the concentrations of trace biomarkers were estimated using the partial least squares regression model applied to the spectra of pure compounds and the biofluids spectrum. Whole blood and saliva simulates spiked with prepared concentrations of the various biochemical components ranging from 1 ppm to 500 ppm were used for for method development. Then, various optimized machine learning techniques that included independent component analysis (ICA), multidimensional scaling (MDS), partial least square discriminant analysis (PLS-DA), kernel density estimators, support vector machines (SVM), and backpropagation neural networks (BPNN) were utilized to analyze and classify the blood and saliva trace biomarkers' Raman spectra from healthy and diseased subjects.

Results using pairwise comparison of mean intensity (peak intensity ratios) and multivariate statistical techniques disclosed that biochemical changes of proteins, lipids, and nucleic acid components can be associated with prostate cancer, breast cancer, and leukemia progression. Four prominent regions: cytosine / guanine ($566 \pm 0.70 \text{ cm}^{-1}$), glycerol (630 cm^{-1}), saccharides ($1370 \pm 0.86 \text{ cm}^{-1}$), tryptophan ($1618 \pm 1.73 \text{ cm}^{-1}$); and six subtle regions: phospholipids (1076 cm^{-1}), amide III ($1232, 1234 \text{ cm}^{-1}$), amide III ($1276, 1278 \text{ cm}^{-1}$),

phospholipids / nucleic acids (1330, 1333 cm^{-1}), lipids (1434, 1442 cm^{-1}), amide II (1471, 1479 cm^{-1}) were identified, which can be regarded as useful biomarkers for prostate cancer diagnosis. Six spectral bands were determined: glycerol (589 cm^{-1}), tryptophan / phosphatidylinositol (594 cm^{-1}), glutamate / tryptophan (630 cm^{-1}), glutamate (1626 cm^{-1}), glycine / valine (1630 cm^{-1}), and amide I / β -carotene (1638 cm^{-1}) which can be regarded as new biomarkers of breast cancer in the blood-based breast cancer spectroscopy.

The fitting model revealed that trace proteins, nucleic acids, and lipid biochemicals in blood and saliva increased with breast malignancy, whereas amounts of glycogen decreased with progression of breast malignancy. For blood samples, the determined concentrations of proteins, saccharides, amino acids, nucleic acids and lipids components in diseased patients were in the range of 237.82-384.96 ppm, 36.4-84.3 ppm, 14.31-83.69 ppm, 66.4-96.8 ppm, and 71.95-297 ppm, respectively, whereas respective concentrations in control samples were 233.86 ppm, 73.7 ppm, 10.48 ppm, 62.1 ppm, and 18-190 ppm. For saliva samples, concentrations of 62.5-126.3 ppm, 11.5-33.9 ppm, 4.90-20.6 ppm, 7.60-9.16 ppm, and 359.6 ppm representing trace proteins, saccharides, amino acids, nucleic acids and lipids in diseased patients were obtained. The respective concentrations in control samples were 27.7 ppm, 33.9 ppm, 2.17-3.66 ppm, 7.35 ppm, and 43.9-145.2 ppm.

The quantitative analysis based on the selected trace biomarker regions suggested that biochemical changes of proteins and membranous lipids increased with leukemia malignancy whereas biochemical changes of nucleic acids, glycogen, and non-membranous lipids decreased with leukemia malignancy. For blood samples, the determined concentrations of proteins, saccharides, amino acids, nucleic acids and lipids components in diseased patients were 6.14 ppm, 2.8 ppm, 1.89-11.1 ppm, 32.25 ppm, and 2.21-3.9135 ppm, respectively, whereas respective concentrations in control samples were 4.04 ppm, 2.72 ppm, 2.29-14.7 ppm, 15.61 ppm, and 4.32-7.1565 ppm. For saliva samples, concentrations of 8.737 ppm, 7.82 ppm, 15.88-17.80 ppm, 5.077 ppm, 0.282-3.645 ppm representing trace proteins, saccharides, amino acids, nucleic acids and lipids in diseased patients were obtained. The respective concentrations in control samples were 11.39 ppm, 14.90 ppm, 1.72-5.04 ppm, 1.069 ppm, and 1.81-4.769 ppm.

The cross-validated models utilized to analyze and classify the blood and saliva Raman spectra from healthy subjects, breast tumor patients, and leukemia patients yielded diagnostic sensitivities of 46% to 100%, as well as specificities of 71% to 100%. Although the number of samples involved in this study were few, the results demonstrate that analysis of Raman spectra of

blood and saliva using optimized machine learning diagnostic algorithms has great potential for the noninvasive and label-free detection of breast cancer and leukemia.

Table of contents

| | |
|--|-----------|
| Declaration | ii |
| Dedication | iii |
| Acknowledgements | iv |
| Abstract | vi |
| Table of Contents | ix |
| List of Abbreviations | xiii |
| List of Figures | xv |
| List of Tables | xx |
| Chapter 1 Introduction | 1 |
| 1.1 Background | 1 |
| 1.1.1 Raman spectroscopy | 2 |
| 1.1.2 Machine learning techniques | 5 |
| 1.2 Problem Statement | 8 |
| 1.3 Research objectives | 8 |
| 1.3.1 General objective | 8 |
| 1.3.1.1 Specific objectives | 8 |
| 1.4 Justification and significance of the study | 9 |
| 1.5 Scope and limitations of the study | 9 |
| 1.6 Study hypothesis | 10 |
| Chapter 2 Literature Review | 11 |
| 2.1 Introduction..... | 11 |
| 2.2 Breast cancer | 14 |
| 2.3 Leukemia | 18 |
| 2.4 Machine learning-fluid biomarkers detection approach for cancer diagnosis..... | 19 |
| Chapter 3 Principles of Quantitative Raman spectroscopy | 26 |
| 3.1 Basics of Raman spectroscopy | 26 |
| 3.2 Theory of Raman spectroscopy | 28 |
| 3.3 Raman spectrometric instrumentation | 31 |
| 3.4 Calibration regression for quantitative Raman spectral analysis | 33 |
| 3.5 Multivariate machine learning techniques for Raman spectral analysis | 35 |

| | |
|--|-----------|
| 3.5.1 Principal component analysis (PCA) | 35 |
| 3.5.2 Linear discriminant analysis (LDA) | 36 |
| 3.5.3 Independent component analysis (ICA) | 36 |
| 3.5.4 Multidimensional scaling (MDS) | 37 |
| 3.5.5 Support vector machine (SVM) | 37 |
| 3.5.6 Back propagation neural network (BPNN) | 38 |
| Chapter 4 Materials and Methods | 37 |
| 4.1 System configuration and optimization | 39 |
| 4.1.1 STR Raman spectrometer system | 39 |
| 4.1.1.1 STR Raman spectrometer system power loss characteristics | 39 |
| 4.1.1.2 Stability of laser output characteristics | 42 |
| 4.1.1.3 Wavelength stability of laser | 43 |
| 4.1.1.4 Choice of optimal substrates for biological samples measurements | 43 |
| 4.1.1.5 Effects of ambient light on measurements | 44 |
| 4.1.1.6 The effect of power density and exposure times on signal-to-noise ratio (<i>SNR</i>) | 46 |
| 4.2 Cell culture: <i>in vitro</i> application in cancer research | 48 |
| 4.2.1 PC3 and PNT1a cells preparation | 49 |
| 4.3 Blood and saliva samples collection | 50 |
| 4.3.1 Inclusion and exclusion criteria | 52 |
| 4.3.1.1 Breast cancer | 52 |
| 4.3.1.2 Leukemia | 52 |
| 4.4 Sample preparation | 53 |
| 4.4.1 Whole blood and saliva biofluid samples | 53 |
| 4.4.2 Whole blood and saliva simulates | 53 |
| 4.4.3 Calibration set design for biochemical components formulation | 53 |
| 4.5 Raman spectral data collection | 54 |
| 4.6 Raman spectral analysis | 57 |
| 4.6.1 Determination of prominent Raman bands for cancer diagnostics | 57 |
| 4.6.2 Determination of trace Raman bands for cancer diagnostics | 57 |
| 4.6.3 Quantitative Raman spectral analysis using partial least-squares regression | 60 |
| 4.6.4 Multivariate statistical analysis of trace biomarkers alterations | 62 |
| Chapter 5 Results and Discussion | 66 |
| 5.1 Raman spectroscopy characterization of PC3 and PNT1a cells | 66 |

| | |
|--|-----|
| 5.1.1 Analysis of prominent biochemical alterations in prostatic cells | 66 |
| 5.1.2 Analysis of trace biochemical alterations in prostatic cells | 76 |
| 5.2 Raman spectroscopy characterization of blood and saliva fluids for breast cancer diagnostics | 84 |
| 5.2.1 Raman spectroscopy characterization of blood | 86 |
| 5.2.1.1 Analysis of prominent biochemical alterations in whole blood spectra | 86 |
| 5.2.1.2 Analysis of trace biochemical alterations in whole blood spectra | 91 |
| 5.2.1.3 Quantitative analysis of trace biomarkers in whole blood spectra using partial least-squares regression | 101 |
| 5.2.1.4 Multivariate exploratory analysis of Independent Component Analysis (ICA), Multidimensional Scaling (MDS), and Partial least Square Discriminant analysis (PLS-DA) for breast cancer diagnostics in blood..... | 113 |
| 5.2.1.5 Multivariate exploratory analysis of Support Vector Machine (SVM) and Backpropagation neural network (BPNN) for breast cancer diagnostics in blood... | 123 |
| 5.2.2 Raman spectroscopy characterization of saliva for breast cancer diagnosis | 134 |
| 5.2.2.1 Analysis of prominent biochemical alterations in saliva spectra | 134 |
| 5.2.2.2 Analysis of trace biochemical alterations in saliva spectra | 135 |
| 5.2.2.3 Quantitative analysis of trace biomarkers in saliva spectra using partial least-squares regression | 144 |
| 5.2.2.4 Multivariate exploratory analysis of Independent Component Analysis (ICA), Multidimensional Scaling (MDS), Partial least Square Discriminant Analysis (PLS-DA) and kernel density estimators for breast cancer diagnostics in saliva.... | 155 |
| 5.2.2.5 Multivariate exploratory analysis of Support Vector Machine (SVM) and Backpropagation neural networks (BPNN) for breast cancer diagnostics in saliva | 162 |
| 5.3 Raman spectroscopy characterization of whole blood and saliva fluids for leukemia diagnostics | 171 |
| 5.3.1 Analysis of prominent biochemical alterations in whole blood and saliva spectra | 171 |
| 5.3.2 Analysis of trace biochemical alterations of blood and saliva for leukemia diagnosis | 177 |
| 5.3.3 Quantitative analysis of trace biomarkers in blood and saliva using partial least-squares regression | 179 |
| 5.3.4 Multivariate statistical analysis of blood and saliva spectra for leukemia diagnostics | 186 |

| | |
|--|-----|
| Chapter 6 Conclusions and Recommendations | 190 |
| References | 196 |
| Appendix I Clinical diagnosis of malignant breast tumor patients and healthy subjects | 216 |
| Appendix II Clinical diagnosis of leukemia patients and healthy subjects | 216 |
| Appendix III Principal Component Analysis – Linear Discriminant Analysis (PCA-LDA) | 217 |
| Appendix IV Independent component analysis (FASTICA_version) | 229 |
| Appendix V Partial least squares discriminant analysis (PLS-DA) | 237 |
| Appendix VI Multidimensional Scaling: Euclidean, mahalanobis, minkowski | 251 |
| Appendix VII Potential functions | 253 |
| Appendix VIII Backpropagation neural networks | 264 |
| Appendix IX Support vector machine (SVM) | 279 |

LIST OF ABBREVIATIONS

| | |
|----------|--|
| CCD | Charge coupled device |
| NIR | Near infrared |
| IR | Infrared |
| FTIR | Fourier transform infrared |
| ATR-FTIR | Attenuated total reflection-Fourier transform infrared |
| DNA | Deoxyribonucleic acid |
| RNA | Ribonucleic acid |
| CTC | Circulating tumor cells |
| ALL | Acute lymphocytic |
| CLL | Chronic lymphocytic |
| AML | Acute myeloid |
| CML | Chronic myeloid |
| PCs | Principal components |
| FASTICA | Fast independent component analysis |
| JADE | Joint approximate diagonalization of eigenmatrices |
| KICA | Kernel independent component analysis |
| MF-ICA | Mean-field independent component analysis |
| SNR | Signal-to-noise ratio |
| DMEM | Dulbecco's modified eagle's medium |
| EMEM | Eagles minimum essential medium |
| FBS | Fetal bovine serum |
| DPSS | Diode-pumped solid-state |
| CL | Collimator lens |
| ND | Neutral density filter |
| SND | Shutter neutral density filter |
| BP | Band pass filter |
| B.S. | Beam splitter |
| LPF | Long pass filter |
| TS | Translational stage (XYZ stage) |
| PC | Personal computer |

USB Universal serial bus.
2D Two dimension
MLTs Machine learning techniques
PCA Principal Component Analysis
PLS Partial least squares
PLS-DA Partial least squares discriminant analysis
ICA Independent component analysis
SOM Self-organizing maps
AFN Auto-associative feedforward neural networks
LDA Linear Discriminant Analysis
ANN Artificial neural networks
SVM Support vector machines
W.H.O. World Health Organization
SERS Surface-enhanced Raman spectroscopy
PLS Partial least squares
RMSEP Root mean squared error of prediction
LD Limit of detection
LQ Limit of quantification
RSD Relative standard deviation
MDS Multidimensional scaling
BPNN Backpropagation neural networks
mW Milli watts
LCD Liquid crystal display
SVD Singular value decomposition

LIST OF FIGURES

| | | |
|-----|--|----|
| 3.1 | Diagram of energy levels demonstrating processes of anti-Stokes, Rayleigh, and anti-Stokes Raman scattering | 27 |
| 3.2 | The schematic structure of a Raman instrument | 30 |
| 3.3 | Schematic illustration of a typical Raman spectrometric set up | 32 |
| 4.1 | Schematic diagram of a customized STR Raman system | 40 |
| 4.2 | Plot of laser output power at 785 nm (watts / square meter) and ambient temperature (°C) versus time (hours) | 42 |
| 4.3 | (a) The measured crystalline silicon position at 519 - 520 cm^{-1} against time (in minutes), and (b) mean wavenumber position of crystalline silicon ($519.49 \pm 0.0075 \text{ cm}^{-1}$) | 44 |
| 4.4 | Average spectrums of ordinary glass slides, silver coated glass slide, and calcium fluoride substrates | 45 |
| 4.5 | Spectrum showing the effects of ambient light on spectral curve of calcium fluoride substrate in (a) 200-600 cm^{-1} , and (b) 1200-1600 cm^{-1} regions | 46 |
| 4.6 | Raman spectra of (a) blood, and (b) saliva from a grade 3 breast cancer patient, demonstrating the effect of exposure times on <i>SNR</i> of selected bands, in total exposure times of 60 seconds (i), 90 seconds (ii), and 120 seconds (iii) | 47 |
| 4.7 | Schematic overview of the data processing and machine learning steps explored in this study | 58 |
| 5.1 | Photomicrographs of (a) PC3 and (b) PNT1a monolayer grown cells on calcium fluoride (CaF ₂) substrates | 66 |
| 5.2 | (a) Examples of as-collected raw spectra in PC3 and PNT1a cells, and (b) the mean spectra of cell samples | 68 |
| 5.3 | The difference spectrum between normalized (a) stage 1 (48 hours), (b) stage 2 (72 hours), and (c) stage 3 (96 hours) PC3 and PNT1a spectral datasets | 70 |
| 5.4 | The score plots of PC 1 and PC 2 for (a) stage 1 (48 hours), (b) stage 2 (72 hours), and stage 3 (96 hours) spectral measurements | 74 |
| 5.5 | The second principal component (PC 2) loadings that explain discrimination of PC 3 cell scores and PNT1a scores | 75 |

| | | |
|-------------|---|-----|
| 5.6 | Scree plots showing eigenvalues explained as a function of the number of principal components for (a) stage 1, (b) stage, and (3) stage 3 spectral datasets | 77 |
| 5.7 | The canonical variable distribution plots (a, c, e) showing distinct positions of significant principal components, and the score plots of intermediate and higher-order PCs and the diagnostic line from LDA for (b) stage 1, (d) stage 2, and (f) stage 3 spectral measurements | 79 |
| 5.8 | Loading vectors explaining the distribution of low- order PC (PC 1) scores against the intermediate-order PCs scores (PC 5 for stage 1, PC 8 for stage 2, PC 5 for stage 3) in (a) stage 1 (48 hours), (b) stage 2 (72 hours) and (c) stage 3 (96 hours) spectral datasets | 81 |
| 5.9 | (a) The optical photomicrograph of a dried blood sample with a laser spot (+) indicated (50x magnification), (b) the overall stacked mean spectra (normal: ($n = 23$); diseased: ($n = 20$), (c) Raman alterations in controls ($n = 23$), grade 1 ($n = 3$), grade 2 ($n = 5$), and grade 3 ($n = 10$) diseased samples, and (d) spectra differences between Raman spectra of healthy and diseased samples | 87 |
| 5.10 | The log scree plots that explain scores in (a) grade ₁ , (b) grade ₂ and (c) grade ₃ spectral datasets for blood samples from healthy (normal) and breast cancer patients | 91 |
| 5.11 | Canonical variable distribution showing the low-, intermediate- and high-order principal components for (a) grade ₁ , (b) grade ₂ and (c) grade ₃ spectral datasets | 93 |
| 5.12 | Scatter plots showing distribution of the first principal component (PC1) versus various PCs (2, 3, 7, 12) scores and the diagnostic line for LDA for (a, b) grade ₁ , (c, d) grade ₂ , and (e, f) grade ₃ spectral datasets | 96 |
| 5.13 | Linear discriminant functions showing loading vectors associated with scores discrimination | 97 |
| 5.14 | Linear discriminant functions showing loading vectors associated with scores discrimination due to intermediate and high-order principal components | 99 |
| 5.15 | Mean Raman spectra of the biochemical constituents used in the concentration fit | 103 |
| 5.16 | Regression plots for partial least squares measured versus predicted biochemical concentrations of the basal compounds used in the spectral mode | 104 |
| 5.17 | Plot of the relative contribution of selected basal compounds estimated by the Raman spectral model applied to the spectra of blood samples from the controls, grade 1, grade 2, and grade 3 breast cancer patients | 106 |

| | | |
|-------------|--|-----|
| 5.18 | PLS-DA scatterplots showing differentiation of spectra of healthy samples, grade 1, grade 2, and grade 3 breast cancer samples | 108 |
| 5.19 | The loading vector plots that explain differentiation of (a) controls versus grade 1 breast cancer, (b) controls versus grade 2 breast cancer, and (c) controls versus grade 3 breast cancer | 110 |
| 5.20 | The loading vector plots that explain differentiation of (a) grade 1 versus grade 2 breast cancer, and (b), (c) grade 2 versus grade 3 breast cancer | 110 |
| 5.21 | The ICA eigenvalues for Raman spectra of blood samples from healthy volunteers and (a) grade 1, (b) grade 2, and (c) grade 3 breast cancer patients | 114 |
| 5.22 | The ICA-PLS-DA scatter plots for Raman spectra of blood samples from healthy volunteers and (a) grade 1, (c) grade 2, and (e) grade 3 breast cancer patients | 117 |
| 5.23 | The spectral markers for independent components of Raman spectra of blood samples from healthy volunteers and (a), (b) grade 1, and (c), (f) grade 2 breast cancer patients | 118 |
| 5.24 | The spectral markers for (a-c) independent components of Raman spectra of blood samples from healthy volunteers and grade 3 breast cancer patients | 119 |
| 5.25 | The ICA followed by MDS-PLS-DA scatter plots of Raman spectra of blood samples from healthy volunteers and (a) grade 1, (b) grade 2, and (c) grade 3 breast cancers patients ... | 122 |
| 5.26 | The SVM scatter plots for breast cancer detection of (a, b) grade 1 ($n = 26$), (c, d) grade 2 ($n = 30$), and (e, f) grade 3 breast cancer ($n = 33$) spectral datasets | 124 |
| 5.27 | The scatter plots of BPNN diagnostics model for detection of of (a) grade 1, (b) grade 2, and (c) grade 3 breast cancer | 126 |
| 5.28 | SVM prediction models for breast cancer detection of (a, b) grade 1 cancer, (c, d) grade 2 cancer, and (e, f) grade 3 cancer | 130 |
| 5.29 | BPNN predictor models for (a) grade 1 cancer, (b) grade 2 cancer and (c) grade 3 cancer | 131 |
| 5.30 | (a) Photomicrograph of dried saliva with a laser spot (+) indicated at x50 magnification, and (b) prominent saliva Raman bands amongst the control (normal) and diseased samples | 134 |
| 5.31 | Overall spectra differences between Raman spectra of (a) the control ($n = 23$) and all diseased ($n = 20$) samples, (b) the control ($n = 23$) and grade 1 ($n = 3$) samples, (c) the control ($n = 23$) and grade 2 ($n = 7$) samples, and (d) the control ($n = 23$) and grade 3 ($n = 10$) samples | 136 |

| | | |
|-------------|--|-----|
| 5.32 | The log scree plots that explain overall scores discrimination in (a) grade ₁ , (b) grade ₂ and (c) grade ₃ spectral datasets of saliva samples from healthy volunteers (controls) and breast cancer patients | 138 |
| 5.33 | Canonical variable distribution showing the low- and intermediate- order principal components for (a) grade ₁ , (b) grade ₂ , and (c) grade ₃ saliva spectral datasets | 139 |
| 5.34 | Scatter plots showing distribution of low-order PC (PC1) scores versus (a) PC 17 scores, (b) PC 19 scores, and (c) PC 13 scores | 142 |
| 5.35 | Loading functions explaining scores discrimination of control and diseased saliva spectra | 143 |
| 5.36 | Regression plots for partial least squares measured versus predicted biochemical concentrations of the pure biochemical compounds | 146 |
| 5.37 | PLS-DA scatterplots showing differentiation of (a) controls from grade 1, (b) controls from grade 2, (c) controls from grade 3, (d) grade 1 from grade 2, and (e) grade 2 from grade 3 breast cancers | 150 |
| 5.38 | Loading functions explaining differentiation of (a) controls from grade 1 breast cancer, and (b) controls from grade 2 breast cancer | 151 |
| 5.39 | Loading functions explaining differentiation of (a) controls from grade 3 breast cancer scores, and (b) grade 1 breast cancer from grade 2 breast cancer | 152 |
| 5.40 | Loading functions explaining differentiation of grade 2 breast cancer from grade 3 breast cancer | 153 |
| 5.41 | The eigenvalues for Raman spectra of saliva samples from healthy volunteers and (a) grade 1, (b) grade 2, and (c) grade 3 breast cancer patients | 156 |
| 5.42 | The ICA followed by PLS-DA scatter plots for (a) grade 1, (c) grade 2, and (e) grade 3 spectral datasets of saliva samples from control and breast cancer patients | 158 |
| 5.43 | The spectral markers for independent components of Raman spectra of saliva samples from healthy volunteers and (a, b) grade 1, and (c, d) grade 2 breast cancer patients | 159 |
| 5.44 | The spectral markers for independent components of Raman spectra of saliva samples from healthy volunteers and (a, d) grade 3 breast cancer patients | 160 |
| 5.45 | The diagnostic results of ICA followed by multidimensional scaling and kernel density estimators for (a) grade 1 breast cancer, (b) grade 2 breast cancer, and (c) grade 3 breast cancer | 163 |
| 5.46 | The SVM training models for breast cancer detection for (a, b) grade 1 breast cancer, (c, d) grade 2 breast cancer, and (e, f) grade 3 breast cancer | 166 |

| | | |
|-------------|---|-----|
| 5.47 | SVM prediction models of breast cancer detection for (a, b) grade 1 breast cancer, (c, d) grade 2 breast cancer, and (e, f) grade 3 breast cancer | 167 |
| 5.48 | The BPNN training (a, c, e) and predictor (b, d, f) models for detecting (a, b) grade 1 breast cancer, (c, d) grade 2 breast cancer, and (e, f) grade 3 breast cancer | 169 |
| 5.49 | Mean normalized spectra of (a) whole blood and (b) saliva for healthy volunteers / controls ($n = 12$) and leukemia ($n = 9$) patients | 171 |
| 5.50 | The difference spectrum for (a) blood spectra, and (b) saliva spectra from healthy volunteers (controls) and leukemia patients | 172 |
| 5.51 | The scree plots showing the number of optimal number of principal components (PCs) for (a) blood spectra and (b) saliva spectra, and the canonical variable distributions of the first twenty principal components for (c) blood spectra and (d) saliva spectra | 175 |
| 5.52 | Scatter plot of the linear discriminant analysis demonstrating the clustering of (a) whole blood spectra and (b) saliva spectra of healthy volunteers and leukemia patients | 176 |
| 5.53 | Scores plot for the higher order components (PC 5, 6) versus the first principal component (PC 1) for (a) whole blood and (b) saliva Raman spectra (red leukemia samples, blue controls), and (c, d) loading vectors for PC 5 and PC 6 | 178 |
| 5.54 | Regression plots for partial least squares measured versus predicted biochemical concentrations of the basal compounds used in the spectral model, based on the spectra profiles of whole blood samples | 180 |
| 5.55 | Regression plots for partial least squares measured versus predicted biochemical concentrations of the basal compounds used in the spectral model, based on the spectra profiles of saliva samples | 181 |
| 5.56 | Scatter plots of RBF-SVM and BPNN diagnostic models demonstrating clustering of Raman spectra of blood samples from healthy volunteers and leukemia patients | 187 |
| 5.59 | Scatter plots of RBF-SVM and BPNN diagnostic models demonstrating clustering of Raman spectra of saliva samples from healthy volunteers and leukemia patients | 187 |

LIST OF TABLES

| | | |
|------|---|---------|
| 1.1 | Incidence of cancers in Nairobi County during 2004–2008 period | 12 |
| 4.1 | The power losses in the STR Raman system's main components | 41 |
| 4.2 | The <i>SNR</i> values of saliva and blood Raman peaks measured at different exposure times | 47 |
| 4.3 | Typical concentration ranges of biochemical components in whole blood and saliva in a human body | 55 |
| 4.4 | Calibration set design for biochemical components formulation in whole blood and saliva | 56 |
| 5.1 | Raman band assignments of PC3 and PNT1a prostatic cells | 69 |
| 5.2 | Comparison of peak intensity ratios between malignant (PC3) and normal cells (PNT1a) at $566 \pm 0.70 \text{ cm}^{-1}$, 630 cm^{-1} , $1370 \pm 0.86 \text{ cm}^{-1}$, and $1618 \pm 1.73 \text{ cm}^{-1}$ spectral regions | 72 |
| 5.3 | Categorization of principal components (PCs) based on their cumulative percentage of total variation and the variance sizes: low-, intermediate-, and high-order PCs | 76 |
| 5.4 | The Student <i>t</i> -test (<i>p</i> -values), and effect sizes (Cohen- <i>d</i> , Pearson's correlation coefficients (<i>r</i>)) showing the relationship between the principal component scores of normal (PNT1a) and malignant (PC3) cells | 78 |
| 5.5 | The Raman bands (loading vectors) that explain natural groupings of PNT1a and PC3 scores using the fifth principal component (PC 5 (0.03%)) for stage 1, eighth principal component (PC 8 (0.02%)) for stage 2 and fifth principal component (PC 5 (0.01%)) for stage 3 spectral datasets | 83 |
| 5.6 | Comparison of peak intensity ratios between malignant (PC3) and normal cells (PNT1a) based on the subtle Raman peaks at 1076 cm^{-1} and 1232 cm^{-1} | 84 |
| 5.7 | Raman band assignments of healthy people and breast cancer patients | 89-90 |
| 5.8 | Categorization of principal components (PCs) based on the cumulative percentage of total variation and the size of variances | 92 |
| 5.9 | The <i>t</i> -test (<i>p</i> -values), and effect sizes ((Cohen- <i>d</i> , Pearson's correlation coefficients (<i>r</i>)) showing relationship between the principal component scores of control and diseased blood samples..... | 94 |
| 5.10 | Raman band assignments of alterations in blood samples of healthy and breast cancer patients | 100-101 |

| | | |
|-------------|--|---------|
| 5.11 | Detection limits of biochemical components for Raman analysis of simulate blood fluid | 103 |
| 5.12 | Comparison of biochemical components concentrations in a whole blood simulate reference solution and the results obtained from PLS regression | 105 |
| 5.13 | Relative amounts of biochemical components in blood samples of healthy and breast cancer patients-based on the determined trace biomarker alterations | 105 |
| 5.14 | Diagnostic results of PLS-DA on the Raman spectra of whole blood from healthy volunteers (controls) and breast cancer patients | 112 |
| 5.15 | Selected dimensions (eigenvalues) and respective explained total variances for ICA by Maximum Likelihood (ML) fast fixed-point estimation on Raman spectra of blood samples from healthy volunteers (control) and breast cancer patients | 115 |
| 5.16 | Comparison of chemometric results on subtle spectral markers using PLS-DA and ICA- PLS-DA techniques | 116 |
| 5.17 | Diagnostic results of ICA followed by PLS-DA on the Raman spectra of blood from healthy volunteers (controls) and breast cancer patients | 120 |
| 5.18 | Diagnostic results of ICA followed by MDS and PLS-DA on the Raman spectra of blood from healthy volunteers (controls) and breast cancer patients | 121 |
| 5.19 | SVM models characteristics for diagnostic analysis on the Raman spectra of whole blood from healthy volunteers (controls) and breast cancer patients..... | 125 |
| 5.20 | Diagnostic results of linear and RBF-SVM models on the Raman spectra of whole blood from healthy volunteers (controls) and breast cancer patients | 125 |
| 5.21 | Diagnostic results of BPNN diagnostic model on the Raman spectra of whole blood from healthy (controls) and breast cancer stricken patients | 127 |
| 5.22 | Diagnostic results of linear and RBF SVM predictor models on the Raman spectra of whole blood from healthy (controls) and breast cancer patients | 129 |
| 5.23 | Diagnostic results of BPNN predictor model on the Raman spectra of whole blood from healthy volunteers (controls) and breast cancer patients | 131 |
| 5.24 | Raman band assignments of saliva from healthy volunteers and breast cancer patients | 137-138 |
| 5.25 | Categorization of PCs based on the cumulative percentage of total variation and the size of variances: Low-order PCs (<90% of the cumulative variance and >1.0 average eigenvalue), Intermediate-order PCs (between 90% and 95% of the cumulative variance), Higher-order PCs (>95% of the cumulative variance) | 138 |

| | | |
|-------------|--|-----|
| 5.26 | The statistical values (<i>t</i> -test (<i>p</i> -values) and effect sizes ((Cohen- <i>d</i> , Pearson's correlation coefficients (<i>r</i>)) showing relationship between the principal component scores of control and diseased saliva samples | 141 |
| 5.27 | Comparison of biochemical components concentrations in a saliva simulate reference solution and the results obtained from PLS regression | 145 |
| 5.28 | Detection limits (mg/ml) of biochemical components for Raman analysis of simulate saliva | 145 |
| 5.29 | Relative amounts of biochemical components in saliva of control and breast cancer patients in fingerprint (500-1800 cm ⁻¹) and the selected subtle band spectral regions | 147 |
| 5.30 | Diagnostic results of PLS-DA on the Raman spectra of saliva from healthy volunteers (controls) and breast cancer patients | 153 |
| 5.31 | Selected dimensions (eigenvalues) and explained total variances for ICA by Maximum Likelihood (ML) fast fixed-point estimation on Raman spectra of saliva samples from healthy volunteers (control) and breast cancer patients | 155 |
| 5.32 | Diagnostic results of ICA followed by PLS-DA on the Raman spectra of saliva from healthy (controls) and breast cancer patients | 161 |
| 5.33 | Diagnostic results of ICA followed by MDS and kernel density estimators (potential function analysis) on the Raman spectra of saliva from healthy volunteers (controls) and breast cancer patients | 162 |
| 5.34 | SVM models characteristics for diagnostic analysis on the Raman spectra of saliva samples from healthy (controls) and breast cancer patients | 164 |
| 5.35 | Diagnostic results of linear-SVM and RBF-SVM models on the Raman spectra of saliva from healthy volunteers (controls) and breast cancer patients | 165 |
| 5.36 | Diagnostic results of linear-SVM and RBF-SVM predictor models on the Raman spectra of saliva samples from healthy (controls) and breast cancer stricken patients | 168 |
| 5.37 | Diagnostic results of BPNN training model on the Raman spectra of saliva samples from healthy volunteers (controls) and breast cancer patients | 170 |
| 5.38 | Diagnostic results of BPNN predictor model on the Raman spectra of saliva samples from healthy volunteers (controls) and breast cancer patients | 170 |
| 5.39 | Detection limits of biochemical components for Raman analysis of simulate blood fluid | 182 |
| 5.40 | Detection limits of biochemical components for Raman analysis of simulate saliva fluid | 182 |

| | | |
|-------------|---|-----|
| 5.41 | Comparison of biochemical components concentrations in a whole blood simulate reference solution and the results obtained from PLS regression | 183 |
| 5.42 | Comparison of biochemical components concentrations in a standard saliva simulate and the results obtained from PLS regression | 183 |
| 5.43 | Estimated amounts of biochemical components in whole blood of normal (control) and leukemia patients- based on the fingerprint (500-1800 cm^{-1}) and the subtle bands spectral regions | 184 |
| 5.44 | Estimated amounts of biochemical components in saliva of normal (control) and leukemia patients- based on the fingerprint (500-1800 cm^{-1}) and the subtle bands spectral regions | 184 |
| 5.45 | Diagnostic results of RBF-SVM predictor model on the Raman spectra of blood samples from healthy volunteers (controls) and leukemia patients | 188 |
| 5.46 | Diagnostic results of BPNN predictor model on the Raman spectra of blood samples from healthy volunteers (controls) and leukemia patients | 188 |
| 5.47 | Diagnostic results of RBF-SVM predictor model on the Raman spectra of saliva samples from healthy volunteers (controls) and leukemia patients | 188 |
| 5.48 | Diagnostic results of BPNN predictor model on the Raman spectra of saliva samples from healthy volunteers (controls) and leukemia patients | 188 |

Chapter 1 Introduction

1.1. Background

One of the most serious threats to humanity is cancer, which is one of the most well-known human diseases. Cancer is a disorder marked by the growth of abnormal cells (American Cancer Society, 2017). The GLOBOCAN database indicated 12.7 million and 18.1 new cancer cases occurred in 2008 and 2018, respectively whereas 7.6 million and 9.6 million cancer deaths occurred in 2008 and 2018, respectively (Ferlay *et al.*, 2010; Freddie *et al.*, 2018). By 2030, there could be twenty two million new cancer cases and thirteen million cancer deaths due to the acceptance of risky lifestyles (American Cancer Society, 2015). According to a previous study that determined the incidences of cancers in Nairobi population between 2004-2008, prostate and breast cancers were the most common among men and women respectively, with age standardized incidence rates (ASR) of 40.6 / 100,000 and 51.7 / 100, 000 respectively (Korir *et al.*, 2015).

Traditionally, histopathological examination of biopsy samples (Ci *et al.*, 1999) and imaging techniques (Parawira, 2009; Beata *et al.*, 2012) are generally utilized for breast cancer diagnosis. Similarly, histopathological analysis of blood specimen (complete blood picture test) is the most preferred diagnostic method for leukemia diagnosis. Such diagnostic processes are subjective, time consuming, and costly (Ci *et al.*, 1999; Parawira, 2009; Beata *et al.*, 2012). Furthermore, histopathological processes involve conventional excisional biopsy procedures which could be potentially hazardous due to their invasive nature.

In the recent past, non – invasive real time optical diagnostic (spectroscopic) techniques have been adopted as better alternatives in clinical medicine. The spectroscopy techniques are based on interaction of materials (e.g., biological samples) with electromagnetic radiation via processes such as absorption, transmission, reflection and scattering (Fotakis *et al.*, 2007). Regarding vibrational spectroscopy e.g., Raman and infrared spectroscopy, the sample molecules are excited into their vibrational states when radiation interacts with biological materials, thereby generating fingerprints of active biomolecule vibrations (e.g., nucleic acids, proteins, lipids, and carbohydrates) of tissues and cells. The obtained fingerprints may be used to explain the biochemical and morphological changes during disease progression (e.g., in cancer development).

1.1.1 Raman spectroscopy

Raman spectroscopy is a nondestructive analytical vibrational spectroscopy technique that reveal fingerprints of active biomolecular vibrations from biological tissues and cells. Raman spectroscopy utilizes scattering (Stokes and anti-Stokes scattering) to excite targeted molecules into the vibrational excited states (Fotakis *et al.*, 2007; Matthäus *et al.*, 2008). Photon energy loss or gain may be used to image Stokes and anti-Stokes scattering, respectively. The scattered radiation is dispersed in a spectrometer, recorded and analyzed. The observed Raman bands are typically narrow, relatively easy to resolve, and exhibit specific vibrational modes that reveal the studied sample materials' fingerprints / associated signatures e.g., biomarkers (Matthäus *et al.*, 2008). Furthermore, since molecular vibrations are highly influenced by a molecule's conformation and its chemical environment (Schie, 2013), spectral analysis may help distinguish active molecular bands and assess the impact of the microenvironment on studied samples, such as biological cells. Apart from being a fast and objective technique (Matthäus *et al.*, 2008), Raman spectroscopy also has the following advantages: (i) is far less intrusive, (ii) reagent free in most cases, (iii) has a lot of depth and spatial resolution ($\leq 1\mu\text{m}$), and (iv) water bands have no impact on Raman spectra measurements (Chowdary *et al.*, 2006; Talari *et al.*, 2015). The latter can be attributed to phenomenon of strongest bands associated with water being found in the high wavenumber region, meaning the water bands have little effect on most Raman spectra in rich fingerprint region ($500\text{-}1800\text{ cm}^{-1}$) associated with biological samples (Shipp *et al.*, 2017).

A major drawback with Raman spectroscopy (especially on biological samples) is its weak signals being dominated by broadband fluorescence emissions. This can be attributed to the low percentage of incident photons ($\approx 0.001\%$) that produces inelastic Raman signal suitable for Raman measurements (Matthäus *et al.*, 2008). For instance, though both fluorescence and Raman spectroscopy are based on vibronic effects, Raman scattering is at least 6 orders of magnitude weaker than fluorescence (Matthäus *et al.*, 2008). Nevertheless, Raman spectroscopy (including micro-spectroscopy) possess key features that make it attractive to scientists (Fotakis *et al.*, 2007): The technique is extremely precise due to the peculiar fingerprint existence of the Raman spectrum. Furthermore, Raman microscopes have excellent spatial resolution, which aids in the study of small features, such as cellular analysis. Raman spectroscopy is also nondestructive and can be done in situ, reducing the amount of time spent sampling. Other encouraging advances include the availability of advanced fiber-optic Raman probes for remote sample analysis, the availability of Raman databases for identifying unknown sample components, and the ease with

which Raman setup systems can be used. For instance, Raman microscopes are easy to use and are easily coupled to one's choice of excitation sources, spectrographs and charge-coupled devices (CCD). With above advantages, coupled with utilization of near infrared (NIR) laser excitation (that minimize fluorescence effects) and high throughput dispersive spectrometers, Raman spectroscopy has consequently previously demonstrated its potential for biomedical diagnosis in various malignancy (Chowdary *et al.*, 2006; Rehman *et al.*, 2010; Shafer-peltier *et al.*, 2002).

Raman microspectroscopy was utilized in understanding gross biochemical differences in breast cancer (Chowdary *et al.*, 2006; Rehman *et al.*, 2010; Shafer-peltier *et al.*, 2002) and leukemia (Erukhimovitch *et al.*, 2006; MartinEspinoza *et al.*, 2008; Babrah *et al.*, 2007) progression where various findings regarding biochemical changes in studied samples have been reported. In these studies, biochemical changes majorly attributed to proteins, lipids, nucleic acid components can be associated with onset and progression of breast cancer and / or leukemia progression. However, in spite of tissues being regarded as a gold standard technique for breast cancer diagnosis, the method is costly, invasive; meaning generally uncomfortable to patients, and inconvenient (from a scheduling perspective) (Jr *et al.*, 2014). Similarly, although use of blood fluid as a gold standard method for leukemia diagnosis is well established, little effort has been made to use other body fluids as an alternative technique for leukemia diagnosis, e.g., the saliva-based Raman spectroscopy.

The detection of cancers after they have advanced to the point of being metastatic and drug resistant has proved to be a difficult issue. In order to effectively treat and manage cancer, early detection and timely diagnosis is crucial. This necessitates need for studying biomarkers in the cancer patient's' biofluids (fluid biomarkers) with aim of obtaining supplementary information that could aid early detection of the cancer (Martin *et al.*, 2010). Fluid biomarkers are components in patient's fluids, that would reveal presence of cancer e.g., macromolecules that originate from tumor cells (lipids, proteins, RNA, microRNA, DNA) and circulating cells (circulating tumor cells (CTC), immune cells, stromal cells, endothelial cells) (Martin *et al.*, 2010; Kaczor-Urbanowicz *et al.*, 2017). The utility of body fluids, example, blood and saliva as alternative samples for cancer diagnosis is still a developing field of research interest. Utility of saliva is advantageous because of its noninvasive safe collection, easy to collect, high –speed sampling, heterogeneity in diagnosis, provision of real time information, easy to transport, and convenient storage for later analysis (Pfaffe *et al.*, 2011; Zhang *et al.*, 2016). Similarly, blood is easy to sample and prepare for further analysis (Khanmohammadi *et al.*, 2010). Furthermore, blood and its constituents tend

to be the most convenient for biomarker detection due to their widespread availability, well-established sample collection) protocols, and the ability to replicate the test as frequently as needed to track disease development or treatment response (Baker *et al.*, 2016).

In terms of the utility of blood and saliva for cancer diagnostics, previous histochemistry studies have shown that two major proteins, CA15-3 and c-erB2, are the most important biomarkers for breast cancer diagnosis (Malamud *et al.*, 2011; Singh *et al.*, 2014; Tabak *et al.*, 2001; Agha-Hosseini *et al.*, 2009). In addition, one study found that the HSP90A protein could be used as a biomarker for the metastatic stage of breast cancer (Kazarian *et al.*, 2017). These findings confirm that biochemical changes in serum and salivary proteins can be used as prognostic markers for diagnosis of breast cancer. Previously, biochemical elements of salivary proteins have been used in Raman spectroscopy for breast cancer diagnosis (Feng *et al.*, 2015; Wu *et al.*, 2015), where Raman peaks corresponding to amide regions were pronounced in malignant samples. Previous serum-based Raman studies (Nargis *et al.*, 2019; Pichardo-Molina *et al.*, 2007; Bilal *et al.*, 2017; Vargas-Obieta *et al.*, 2016) have shown that Raman spectral differences in healthy and diseased breast cancer samples can be mainly attributed to DNA, proteins, and lipids alterations. Besides, these findings (Nargis *et al.*, 2019; Pichardo-Molina *et al.*, 2007; Bilal *et al.*, 2017; Vargas-Obieta *et al.*, 2016) suggest that biochemical changes due to serum proteins are predominant during breast malignancy.

In the case of leukemia, enzyme-linked immunosorbent assay methods on saliva showed that salivary DNA and RNA biomarkers could be reliably identified in leukemic patients (Rasi *et al.*, 2011; Chen *et al.*, 2014), implying that nucleic base conformational changes, such as adenine, cytosine, thymine, and uracil, could be useful for leukemia diagnosis. Furthermore, based on several genome and transcriptome studies, a systematic analysis of known elevated circulating miRNAs associated with chronic lymphocytic leukemia indicated salivary miRNAs (95, 29a, 222, 20a, 150, 451, 135a, 486-5p) may be associated with the presence of leukemia (Allegra *et al.*, 2012). Furthermore, salivary 92 miRNA levels have been found to rise, especially in patients with acute myeloid leukemia (Allegra *et al.*, 2012). These results support previous findings (Rasi *et al.*, 2011; Chen *et al.*, 2014), which show that biochemical changes in nucleic acid components play a catalytic role in leukemia proliferation.

Imaging spectroscopy, unlike traditional Raman spectroscopy, simultaneously captures spectral and spatial information in studied samples (Dyson *et al.*, 2004; Bearman *et al.*, 2003). Raman microspectrometry, which combines Raman spectroscopy and microscopy, is a dependable

technique for obtaining spatially resolved spectroscopic information on minute quantities of microscopic structures within a biological sample (Lasch *et al.*, 1997). With this technique, chemical maps" (CM) or functional group maps (Bearman *et al.*, 2003) can be generated which, in essence, can be directly compared to light microscope observations (Lasch *et al.*, 1997). To that end, Raman spectroscopy and imaging were used to analyze noncancerous and cancerous human breast tissues from the same patient (Brozek-Pluska *et al.*, 2012), with the most important variations observed in regions associated with proteins, carotenoids, and lipids. In other research, Raman microspectrometry was used to discern basal cell carcinoma from non-cancerous tissue, pointing to the possibility of creating an *in vivo* diagnostic technique for tumor boundary demarcation (Nijssen *et al.*, 2002). During the examination and classification of optical absorption properties of bone marrow cells in an acute lymphoblastic leukemia sample, spectral microscopy was also used to obtain a large number of narrow-band images in the broad spectral range of the optical spectrum (Katzilakis *et al.*, 2004). In terms of detecting, identifying, and mapping their spectral absorption properties, the study revealed a statistically significant difference ($p < 0.0001$) between normal lymphocytes and lymphoblasts.

The lack of quantification of biochemical changes occurring during cancer development is a disadvantage of blood-based and saliva-based leukemia and breast Raman studies. Another difficulty in bio-spectroscopy is that analytes in blood and saliva are found in low concentrations (Pfaffe *et al.*, 2011; Christodoulides *et al.*, 2005). The problem is exacerbated by a general lack of detailed understanding of the information material that may be accessible from infrared spectra of complex biological samples (Lasch *et al.*, 1997), meaning that robust techniques e.g., machine learning techniques (MLTs) would be crucial for data mining. It is our belief the combination of Raman microspectrometry with machine learning techniques (MLTs) can provide high sensitivity, accuracy, precision and speedy non-destructiveness *in-situ* and *in-vitro* diagnostic capabilities. This can be attributed to their ability of mining complex analytical information between observed variables and measurements (Wang *et al.*, 2014; Varmuza *et al.*, 2008).

1.1.2 Machine learning techniques

Machine learning techniques (MLTs) are computational intelligent methods for extracting maximum analytical information from measured data (Wang *et al.*, 2014). Thus, MLTs are capable of capturing unknown underlying multivariate relationships between observed variables (Varmuza *et al.*, 2008). Preprocessing, feature extraction, and classification algorithms are some of the

machine learning techniques that can be used to process non-resonant Raman spectra in Raman spectroscopy (Rodionova *et al.*, 2006). Generally, the prominent preprocessing techniques utilized to mitigate fluorescence noise among other non-linearities in the measurements include wavelet transformations, derivative filters, baseline subtraction using polynomial fittings, and Principal Component Analysis (PCA) (Lasch, 2012). The wavelet transformation algorithms are purely frequency domain techniques. Apart from being advantageous in filtering out the lower frequency components (which majorly comprise fluorescence noise) from the measured Raman spectra (Lasch, 2012), wavelet transformations techniques are versatile, efficient and offer multi – resolution decomposition (Wang *et al.*, 2014). Nevertheless, challenges associated with wavelet filtering such as under and / or over filtering and subjectivity in analysis, limit their applications in practical clinical set-up. Instead, the polynomial baseline fitting techniques which offer speed, simplicity, convenience and reliability in preserving the Raman line shapes (Lasch, 2012), are generally preferred in many Raman studies. Apart from fluorescence background removal aspects, some techniques may offer smoothing capability without greatly affecting the resultant spectra. These include Savitzky-Golay algorithms and wavelet transformations (Bilal *et al.*, 2017; Hu *et al.*, 2007).

Feature extraction techniques majorly comprise linear and nonlinear techniques that reduce redundancy in spectra variables, thereby retaining most significant biochemical information within the datasets. The commonly known linear methods include Principal Component Analysis (PCA), Partial Least Squares (PLS), and Independent Component Analysis (ICA). Although PCA is a good way to reduce spectra data, ICA has been shown to perform better, especially in non-Gaussian data, by revealing physically measurable components (independent components) and their concentration profiles (Chung *et al.*, 2005; Yao *et al.*, 2012). Nonlinear methods include self-organizing maps (SOMs), multidimensional scaling (MDS), and auto-associative feedforward neural networks (AFN). Intensive computation is one of the drawbacks of nonlinear methods, and their efficiency is highly dependent on parameter selection and configuration (Simeonova *et al.*, 2010; Kolehmainen *et al.*, 2001).

PCA and PLS are often used along with Linear Discriminant Analysis (LDA) algorithms during spectral analysis. Machine learning techniques of LDA, artificial neural networks (ANN), naïve Bayesian, and support vector machines (SVM) are classification techniques that can be used to discriminate spectral data amongst samples (Bird *et al.*, 2008). Being a low dimensional classifier, LDA require already dimensionally reduced data, and fits well with feature space that is

linearly separable (Li *et al.*, 2012). SVM is used to solve problems of two groups (classes) that require optimal efficiency. Nevertheless, it requires lengthy training time and optimal choice of kernel parameters (Gautam *et al.*, 2015). Unlike the Bayesian classifiers that works well on small data sets, ANN easily handles large multi-class data problems, both linear and non – linear in feature space (Wang *et al.*, 2014). Several reviews extensively detailing various applications of MLTs in extracting spectra information during cancer diagnostics are highlighted elsewhere (Wang *et al.*, 2014; Gautam *et al.*, 2015; Kendall *et al.*, 2009; Chandra *et al.*, 2015).

Raman microspectrometry has extensive datasets that include both spectral and spatial details. Detection and / or collection of suitable spectra for subsequent processing, as well as spectral classification or pixel-unmixing, are all part of spectral image analysis (Bearman *et al.*, 2003; Lasch, 2012). Each pixel is assigned to one or more spectrally specified classes during classification. Classification is equivalent to spectral segmentation. A pixel or object is allocated to a single class using one or more of a number of metrics (Lasch, 2012). If pixels are made up of more than one spectral class, the pixels must be "unmixed," resulting in estimates of the percentages of each class present. Two methods can be used to decide which spectra to use for the classification procedure, i.e., reference spectra may be chosen from obvious image structures or from existing spectral libraries (Bearman *et al.*, 2003). Alternatively, statistical analysis techniques such as principal component analysis (PCA) or clustering methods can be used to derive insightful spectra (Bearman *et al.*, 2003; Lasch, 2012).

It should however be noted that reliability of selected MLTs can be limited by problems of poor data quality, low accuracy (overall classification accuracy, sensitivity, specificity), numerical instability and slower real-time processing of information in a real practical clinical setup. These challenges can be attributed to improper initialization (of kernel parameters), high nonlinearity of datasets, and sensitivity to Hughes phenomenon (Bishop *et al.*, 2006). For instance, a challenge could be determining the least number of wavelengths required to disclose significant details for disease diagnosis. While it may seem obvious that having more spectral data and a higher spectral resolution would improve analytical accuracy, this is not always the case. Many wavelengths are likely to be "uninformative," and including them in the dataset simply adds noise, hence demanding inclusion of a dimension reduction algorithm such as PCA. However, if such techniques are optimized in novel ways, e.g., proper adjustment of kernel parameters, application of enhanced feature extraction approaches, and combining optimized dimensional reduction

techniques with advanced classifiers (e.g., ANN, SVM), it may be possible to enhance their information extraction capability.

1.2 Problem statement

The utility of body fluids (such as saliva, blood and urine) in cancer diagnosis utilizing Raman microspectroscopy is attractive but is still an underdeveloped research front. Extraction and multivariate interpretation of the subtle disease biomarkers from the samples spectra (crucial for early cancer diagnosis) requires the combination of sensitive microspectrometry and robust machine learning techniques to realize rapid and comprehensive sensitive cancer diagnostics.

1.3 Research objectives

1.3.1 General objective

The main objective of this research was to design, test and evaluate novel machine learning techniques for detection and quantification of biomarker occurrences and multivariate alterations in saliva and blood that can point to the onset and progression of leukemia and breast cancers via laser Raman spectral analysis.

1.3.1.1 Specific objectives

- i. To identify and determine the concentrations of trace biomarkers of leukemic and breast cancer in saliva and blood using laser Raman microspectroscopy.
- ii. To correlate the obtained biomarker levels in (i) as well as their alterations in the selected body fluids matrices to cancer presence and severity based on concentration levels of biochemical changes and the band ratios of trace spectral markers.
- iii. To apply robust and hybridized machine learning techniques (higher-order PCA, ICA, MDS, PLS-DA, and kernel density estimators) in the extraction and multivariate exploratory analysis and interpretation of the biomarkers embedded in the measured spectra.
- iv. To develop conceptual diagnostic models to detect and characterize breast and leukemia cancers in their various stages based on the information obtained in (i), (ii) and (iii) above, based on support vector machine (SVM) and artificial neural networks (ANN).

- v. To test the developed diagnostic models for proof of concept, to detect and predict the status of breast and leukemia cancers in clinical liquid biopsies samples.

1.4 Justification and significance of the study

Cancer diseases have risen dramatically worldwide in recent years, putting a strain on many families. Cancer affects people of all ages and socioeconomic backgrounds, with the risk of developing cancer rising with age. In the long term, this has had a negative effect on poverty alleviation and sustainable growth. In regard to developing countries, the rapid increase in the number of cancer cases has increased public health crisis with a critical and direct negative impact on the first three Millennium Development Goals (MDGs) namely; poverty, education, and gender equity. The fact that most cancer patients are diagnosed at an advanced stage, when treatment options are minimal, means that prognoses are poor and fatality rates are high.

Highly sensitive and unique biomarkers are needed for early cancer detection. Biomarkers in biofluids, such as whole blood and saliva, may be particularly useful in detecting the existence of early tumors in the body. Although spectroscopy with the help of machine learning techniques has aided human cancers' diagnostics, a major challenge has been development of robust algorithms that would enhance quick and real time detection of cancers at their early stage of development. The problem is further compounded by a challenge in detecting analyte concentrations of biofluid biomarkers that originate from tumor cells which include, deoxyribonucleic acids (DNA), ribonucleic acids (RNA), micro-ribonucleic acids (μ RNA), circulating tumor cells (CTC), proteins, and lipids. Further, machine learning technique algorithms are generally limited by problems of poor prediction quality / accuracy, numerical instability and slower real-time processing of information in a practical clinical setup. This demands utility of optimized machine learning techniques - hereby referred to as novel machine learning techniques. The combination of Raman microspectroscopy and novel machine learning techniques in analyzing body fluids has the potential to reveal weak cellular biochemical and structural alterations that would point to early cancer biomarkers even when the original spectra signals from the spectroscopic devices are weak and prone to noise interference.

1.5 Scope and limitations of the study

This was a matched case-control study that aimed at designing, testing and evaluating novelized machine learning techniques for detection and quantification of biomarker occurrences

and multivariate alterations in saliva and blood that can point to the onset and progression of leukemia and breast cancers via laser Raman spectral analysis. It was problematic to obtain enough sample sizes of the willing diseased and healthy (control) participants. By calculation (Kasiulevičius *et al.*, 2006), the study aimed at enrolling 150 breast cancer case patients with 1 matched control(s) per case to be able to reject the null hypothesis that the odds ratio equals 1 with probability (power) = 0.9. Similarly, about 250 leukemic cancer case patients with 1 matched control(s) per case were expected to be included in this study to be able to reject the null hypothesis that the odds ratio equals 1 with probability (power) = 0.9. All the type I error probabilities associated with the test of these null hypotheses are based on an uncorrected chi-squared statistic evaluation. In this study, a group of 23 healthy volunteers / controls (age 34-56 years) and 20 malignant patients (age 41-65 years) all-female participated in breast cancer study while a group of 18 healthy volunteers / controls (age 20-45 years) and 9 malignant patients (age 24-72 years) both males and females participated in the leukemia study. Subsequently, the biochemical information gathered from the biofluid samples (blood and saliva) proposed in this study may not be generalizable and conclusive for leukemia and breast cancer diagnostics since the willing participants may not represent the required random sample size.

To evaluate whether intermediate and high-order principal components were potentially useful at detecting trace biochemical changes in biological samples' Raman spectra, suitable breast and leukemic cell lines were needed. The necessary breast and leukemic cell lines were unavailable for the study. Therefore, the metastatic androgen insensitive (PC3) and immortalized normal (PNT1a) human prostate cell lines were chosen for a model tissue Raman spectroscopy analysis.

1.6 Study hypothesis

The working hypothesis of this study is that combination of robust machine learning techniques and Raman microspectroscopy will be able to detect and quantify biomarker occurrences and multivariate alterations in saliva and blood of leukemia and breast cancer patients at high degree of sensitivity and specificity accuracies.

Chapter 2 Literature Review

2.1 Introduction

Compared to conventional methods, utility of body fluids for cancer diagnostics via Raman microspectroscopy is attractive but is still a developing field of research interest due to a number of challenges mainly regarding the sensitivity of detection, extraction and quantitative microanalysis of the subtle biomarkers in the spectra and images particularly at single cell level. By 2008, cancer was a leading cause of deaths globally (American Cancer Society, 2011). Cancer was estimated to be a leading cause of death in 91 of 172 countries by the age of 70 years (Freddie *et al.*, 2018). Its prevalence is further enhanced by adoption of risky modern lifestyles such as physical inactivity / sedentary lifestyle, smoking, poor diet, and reproductive factors (American Cancer Society, 2017).

Utility of biofluids e.g., saliva and blood for cancer diagnostics would be a field of interest for the rarely early diagnosed common cancers in women and in Kenya's population, i.e., breast cancer, and leukemia, respectively. There are already two recent studies available in literature in which cancer incidences in Kenya's population are reported (Korir *et al.*, 2015; Antabe *et al.*, 2020). According to these findings, breast cancer had the highest age-standardized incidence rates (> 50) in Kenya's female population (Korir *et al.*, 2015), and accounted for 23% of all cancer cases among women in Kenya (Antabe *et al.*, 2020). Elsewhere, recent GLOBOCAN (2018) report showed breast cancer being the topmost frequent cancer amongst females, with new cases at 12.5% growth (Kenya-Globocan, 2018). Similarly, earlier reports showed leukemia had the lowest age standardized incidence rates (< 3) in Kenya's population (Korir *et al.*, 2015). By 2018, statistics showed the 5-year prevalence (all ages) of leukemia in Kenya's population was 7.55 ($n = 3,845$) (Kenya-Globocan, 2018). Nonetheless, 3200 new cancer cases were estimated in children under the age of 18, with leukemia being one of the most common cancers in children, accompanied by brain cancers, Hodgkin's kidney cancer, cancer of the naso-pharynx, and Non-lymphoma (Ministry of Health, Kenya, 2019). According to data from Kenyatta National Hospital, majority of cancers (64%) are diagnosed when already advanced to malignant stages, when cure is difficult to achieve (Ministry of Health, Kenya, 2019). Table 2.1 summarizes the incidence rates of selected cancers in Nairobi County in the years 2004 – 2008 (Korir *et al.*, 2015).

Table 2.1: Incidence* of cancers in Nairobi County during 2004–2008 period (Korir *et al.*, 2015).

| | | | (Incidence per 100, 000) |
|---------------|-------------------------|--|---|
| Gender | Cancer | Number of patients, % of population | Age standardized incidence rates (ASR) |
| Men | Prostate | 606, 15.6% | 40.6 |
| | Lung and trachea | 99, 2.5% | 5.4 |
| | Leukemia | 105, 2.7% | 2.7 |
| | Oesophagus | 333, 8.6% | 15 |
| | Stomach | 243, 6.2% | 11.1 |
| | Colon, rectum, and anus | 296, 7.6% | 12.1 |
| | Lymphoma | 206, 5.3% | 6.7 |
| Women | Breast | 1154, 22.7% | 51.7 |
| | Cervical | 1073, 21.1% | 46.1 |
| | Lung and trachea | 53, 1% | 3.2 |
| | Leukemia | 71, 1.4% | 2.8 |
| | Oesophagus | 248, 4.9% | 14.8 |
| | Stomach | 193, 3.8% | 11.3 |
| | Colon, rectum, and anus | 242, 4.8% | 12.4 |
| | Lymphoma | 142, 2.8% | 7.1 |

* Incidence rate statistics quantify the number of newly diagnosed cancer cases in a specified population over a defined time period, usually expressed per 100,000 people per year (American Cancer Society, 2011).

The routine diagnostic procedures for breast cancer and leukemia include histopathological examination of biopsy samples, fluorescence, optical bioluminescence, ultrasound, magnetic resonance imaging, X-ray mammography, and computed tomography (Ci *et al.*, 1999; Nargis *et al.*, 2019). However, these methods have a range of disadvantages, including the fact that they are frequently subjective, take a long time to yield results, and expensive (Ci *et al.*, 1999; Beata *et al.*, 2012). Low resolution and sensitivity are also problems, and X-ray uses potentially hazardous radiation, which can be detrimental to patients (Nargis *et al.*, 2019). Moreover, conventional excisional biopsy procedure could be potentially hazardous due to its

invasive nature, and pathological grading may be highly subjective (Beata *et al.*, 2012). Besides, these techniques only reveal cancers at already advanced stages (Nargis *et al.*, 2019). This illustrates the importance of developing extremely responsive, real time, less intrusive, and relatively cost-effective cancer diagnostic methods for early early cancer diagnosis.

The relatively recent preference of liquid biopsy (biofluids) to tissue biopsy is revolutionizing the clinical approach towards cancer diagnosis. Biofluids have specific properties that can be assessed and analyzed objectively as markers of natural biologic processes, pathogenic processes, or therapeutic intervention responses- the so-called biomarkers (Baker *et al.*, 2016). Biomarkers in body fluids can be useful in early-stage disease screening, differential diagnosis of the disease with other conditions, disease prognosis independent of treatment, prediction of treatment response, and disease monitoring (Baker *et al.*, 2016). Biomarkers in body fluids include, (i) blood biomarkers, and (ii) biomarkers from other body fluids; depending on tumor location, for example, pleural fluid, tears, bile, sputum, urine, saliva, pancreatic juice, cerebrospinal fluid, and ascitic fluid (Baker *et al.*, 2016). Circulating tumor cells (CTC) and macromolecules (RNA, DNA, μ RNA, proteins, lipids) are examples of biomarkers (Martin *et al.*, 2010). The principle of looking for cancer biomarkers in biofluids before development of disease symptoms seem to be a promising avenue for early cancer detection. This can be partly attributed to benefits of fluid sampling including being generally accepted, easily repeatable, simple to use, non-intrusive, and cost-effective. Furthermore, biomarkers found in bodily fluids enhance ability to detect a wide range of primary tumors and metastases in the body (Martin *et al.*, 2010).

Given that majority of tumours are vascularized, the cancer biomarkers can be shed into the blood- stream (Baker *et al.*, 2016). Furthermore, blood and its constituents tend to be the most convenient for biomarker detection due to their widespread availability, well-established sample collection) protocols, and the ability to replicate the test as frequently as needed to track disease development or treatment response (Baker *et al.*, 2016). Utilization of saliva for disease diagnosis is relatively practical since its collection is noninvasive, easy to collect, transport, store, and safer for handling unlike other body fluids such as blood (Pfaffe *et al.*, 2011). Besides, it requires minimal sample preparation (Emekli-Altufran *et al.*, 2008). Further, many compounds found in blood e.g. growth factors, hormones, antibodies, and enzymes are also found in saliva, hence providing an alternative mechanism of studying emotional, hormonal, nutritional, and metabolic variations in human. Various processes for example, passive diffusion, ultrafiltration, and active transportation may move these compounds from the bloodstream into the salivary glands (Pfaffe

et al., 2011). Consequently saliva as a biofluid has been widely utilized in diagnosing various diseases such as cardiovascular (Meurman, 2010), renal (Walt *et al.*, 2007), diabetes (Rao *et al.*, 2009), autoimmune disorders (Hu *et al.*, 2007), and systemic malignancies (Farnaud *et al.*, 2010; Streckfus *et al.*, 2004; Streckfus *et al.*, 2000).

Biofluids, like other cells and tissues in human body, have vibrational spectra with distinct bands that indicate their biomolecular composition (Baker *et al.*, 2016). The quest for specific disease markers in biofluids using photonic approaches, primarily vibrational spectroscopic approaches, has recently led to emergence of a new field in biomedical sciences - the biospectroscopy. This field has been extensively utilized in breast cancer and leukemia diagnostics as described in section 2.2 and section 2.3.

2.2 Breast cancer

Raman microspectroscopy has already been widely employed in previous breast malignancy diagnosis. For instance, a previous study suggested the ratio of intensities of the bands of total amounts of protein (I_{3473} / I_{3005}), lipids (I_{2924} / I_{2853}), amide I (I_{1655} / I_{1549}), DNA (I_{1236} / I_{1080}), collagen (I_{1204} / I_{1655}), and carbohydrates (I_{1055} / I_{1467}) contents indicated the proteins, amide I, DNA activities, and carbohydrate levels increased with breast malignancy, while relative number of methyl groups (lipids) and collagen contents decreased with breast malignancy (Venkatachalam *et al.*, 2008). In this study, the band intensity ratios at (I_{3473} / I_{3005}), (I_{2924} / I_{2853}), (I_{1655} / I_{1549}), (I_{1080} / I_{1236}), (I_{1204} / I_{1657}), and (I_{1055} / I_{1467}) were (cancer-1.51; normal-0.67), (cancer-0.82; normal-0.87), (cancer-2.53; normal-1.74), (cancer-2.11; normal-0.97), (cancer-0.56; normal-1.9), and (cancer-1.33; normal-0.69), respectively. In contrast, another study indicated amounts of collagen, fatty contents, nuclear – to – cytoplasm ratios (nucleic acids) increased in all cancerous breast tissues, although the relative increment of collagen was highly pronounced in samples undergoing fibrocystic changes (Haka *et al.*, 2005). Therefore, nucleic acid bases and *de novo* lipogenesis processes may be viewed to play a catalytic role in breast cancer progression (Long *et al.*, 2018). Raman spectroscopic measurements were performed on healthy and diseased breast tissues, where the PCA was utilized for discrimination (Chowdary *et al.*, 2006). The study revealed that spectral profiles of normal tissues indicated pronounced levels of lipids. In contrast, the study revealed that both malignant and benign tissues had more proteins and lower levels of lipids in their spectral profiles. Furthermore, lipids were found in greater abundance in malignant tissues than in benign tissues. Another study (Gonzálezsolís *et al.*, 2011) observed dominance of protein

biochemical alterations where tryptophan, protein and amide III (at 1244 cm^{-1}) biomolecular alterations majorly featured in late stages of breast cancer progression, implying oxidative stress in tissues was a major factor during breast cancer progression (Marinello *et al.*, 2014). Previous reviews demonstrating applicability of IR (including FTIR) Raman spectroscopy in breast cancer diagnostics are provided elsewhere (Wang *et al.*, 2009). From these reviews, it can be concluded that processes of proteins degradation, nucleic acid bases alterations and production of lipids via *de novo* lipogenesis are considered as potential biomarkers for breast cancer diagnostic procedures.

It can be seen from the above studies that breast cancer tissues can be studied to assess the viability of biochemical changes occurring in tissue. The detected changes can be classified into different groups which include normal, fibroadenoma, ductal carcinoma, hyperplasia, and invasive ductal carcinoma (Rehman *et al.*, 2010). Furthermore, because of its sensitivity to changes in the composition and amount of biomolecules in the tissues, Raman spectroscopy may be employed for quantitative and qualitative study of cancerous breast tissues. Despite the capability of Raman spectroscopy in examining biochemical changes in biopsy and hence differentiating the different stages of breast cancer, the utility of tissues as the gold standard for clinical diagnosis of breast cancers has been cited to be occasionally inconvenient (from a scheduling perspective), costly, and mainly invasive thereby uncomfortable to patients (Jr *et al.*, 2014).

The saliva biomolecules (such as DNA, mRNA, microRNA, proteins, metabolites, microbiota) have been previously utilized as biomarkers in exploring systemic malignancies (Malamud *et al.*, 2011; Singh *et al.*, 2014; Tabak, 2001; Agha-Hosseini *et al.*, 2009). Previously, two major proteins i.e., the c-erbB2 and CA15-3 proteins have been the most significant biomarkers for breast cancer diagnosis (Malamud *et al.*, 2011; Singh *et al.*, 2014; Tabak, 2001; Agha-Hosseini *et al.*, 2009). The c-erbB2 is a receptor tyrosine kinase and the CA15-3, is a tumor marker found on cancer cell's surface (Malamud *et al.*, 2011; Singh *et al.*, 2014). These proteins ordinarily shed into the bloodstream and have therefore been regularly used to monitor advanced and metastatic breast cancer cases. For example, scientists have previously discovered that women with breast cancer have higher levels of c-erbB2 and CA15-3 proteins in their saliva than healthy women (Malamud *et al.*, 2011; Singh *et al.*, 2014; Tabak, 2001; Agha-Hosseini *et al.*, 2009). In another serum based study (Duffy, 2006), CA-153 protein was suggested as a possible prognostic marker because of its dramatically increased levels in saliva and serum of breast malignancy patients. This was in line with a recent study that found CA15-3 and HSP90A proteins to be serum biomarkers that may be useful for metastatic breast cancer diagnosis (Kazarian *et al.*, 2017). Elsewhere, the

CA 125 protein biomarker was found to be higher in breast cancer patients' saliva and serum samples than in healthy controls (Mittal *et al.*, 2011). These findings support the idea that serum and salivary proteins play an important role in the detection of breast cancer.

Tumor cells and non-malignant cells, are known to shed DNA into the circulatory system – the so called cell-free DNA (cfDNA) (Jr *et al.*, 2014). Hence, another mechanism for utilizing saliva for cancer diagnostics is comparing (i) DNA - methylation patterns (genome and epigenome studies) and (ii) μ RNA (transcriptome studies) between sick and healthy controls (Zhang *et al.*, 2014). Circulating μ RNA play a key role in cell differentiation and proliferation (Allegra *et al.*, 2012), and have therefore featured in detection of breast malignancy. Previously, research findings have shown that μ RNA is often dysregulated in breast cancer patients as compared to controls patients. For instance, a transcriptomic and proteomic study aimed at discovering and pre-validating biomarkers in saliva for oncological application (Zhang *et al.*, 2014), revealed eight μ RNA biomarkers (S100A8, H6PD, IGF2BP1, CSTA, , GRM1, GRIK1, MDM4, TPT1) and one protein biomarker (CA6) that possessed reliable discriminatory power of 92% accuracy (83% sensitivity and 97% specificity) for classifying between breast cancer patients and controls. In addition, a comprehensive review of known circulating miRNAs by Allegra *et al.*, (2012), has previously observed that the 10b, 34a, 195, let-7, and 223 as the common miRNAs prominently pronounced in breast cancer patients. In addition, other studies have previously observed enhanced levels of DNA bases of proline and valine being associated with advanced stages of breast cancer (Porto-Mascarenhas *et al.*, 2017), further confirming the role of DNA methylation in cancer progression.

Spectroscopically, biochemical components in saliva have traditionally been used for diagnosis of other types of cancer for example, oral cancer (Calado *et al.*, 2019), diabetes (Scott *et al.*, 2010), periodontitis (Gonchukov *et al.*, 2012), and lung cancer (Li *et al.*, 2012). In the case of breast cancer, there are only a few studies in the literature that show salivary proteins and sialic acid components in saliva can be used to diagnose the disease. Biochemical components of salivary proteins have been traditionally employed in diagnosis of breast cancer (Feng *et al.*, 2015; Wu *et al.*, 2015). Significant (Students *t*-test, $p < 0.05$) Raman peaks attributed to amide conformations displayed pronounced Raman signals in breast malignancy, indicating that observed changes in amino acids due to C-N stretching modes of proteins ($1049, 1084 \text{ cm}^{-1}$), amide III (1265 cm^{-1}), amide I (1684 cm^{-1}), CH_3CH_2 wagging mode of collagen (1340 cm^{-1}), phenylalanine (1004 cm^{-1}) increased with malignancy (Movasaghi *et al.*, 2007). Elsewhere, breast cancer patients were

differentiated from benign and healthy subjects using a separate technique based on attenuated total reflection-Fourier transform infrared (ATR-FTIR) spectroscopy in saliva samples (Ferreira *et al.*, 2020). At wavenumber 1041 cm^{-1} , absorbance levels in saliva of breast cancer patients were significantly higher than in benign patients, according to the study. Furthermore, region-of curve (ROC) analysis of 1041 cm^{-1} peak showed that the components in the $1302.9\text{--}1433\text{ cm}^{-1}$ wavenumber range were elevated in saliva of breast cancer patients relative to control and benign patients, suggesting a good accuracy in separating breast cancer from benign and control patients. The 1041 cm^{-1} and $1433\text{--}1302.9\text{ cm}^{-1}$ wavenumber regions can be attributed to biochemical changes due to collagen proteins and CH_2 and / or CH_3 bending, twisting, and wagging modes of collagen proteins and lipids, respectively (Chandra *et al.*, 2015). It can therefore be concluded that saliva proteins could be used as prognostic markers for breast cancer detection.

Different from utility of protein biomarkers for breast cancer diagnostics, SERS on sialic acid in saliva of controls and breast cancer patients was evaluated for breast malignancy diagnosis (Hernández-arteaga *et al.*, 2017). The mean concentration of sialic acid in saliva was found to be predominant (Students *t*-test, $p < 0.05$) in breast cancer patients than among healthy controls, and the test yielded sensitivity and specificity of 94%, and 98%, respectively. The results suggest Raman spectroscopic study of sialic acid components in saliva may be a useful method for breast cancer diagnosis. It is our view that saliva-based Raman studies for breast cancer diagnosis should be extended to include other biomarkers e.g., DNA, lipids, and saccharides, to achieve complementary information for better diagnosis.

The utility of blood biofluid in diagnosis of breast cancer is well established. Previous serum-based Raman studies (Nargis *et al.*, 2019; Pichardo-Molina *et al.*, 2007; Bilal *et al.*, 2017; Vargas-Obieta *et al.*, 2016) have shown that DNA, proteins, and lipids alterations are primarily responsible for Raman spectral variations in healthy and diseased breast cancer patient's samples. First, Pichardo-Molina *et al.*, (2007) found seven band ratios corresponding to proteins, polysaccharides, and phospholipids biomarkers were statistically significant for discriminating between the spectra of control and breast cancer patients. Also, Raman spectral features due to biochemical changes of proteins and DNA were only present in blood samples of breast cancer patients when compared to that of the normal persons (Nargis *et al.*, 2019). Elsewhere, lycopene (1528 cm^{-1}), phosphatidylserine (525 cm^{-1}), quinoid ring (1594 cm^{-1}), calcium oxalate (913 cm^{-1}), and calcium hydroxyapatite (963 cm^{-1}) were identified as biomarkers linked to occurrence of breast cancer whereas Raman shifts assigned for tryptophan (1363 cm^{-1}), proline (930 cm^{-1}), valine (930

cm^{-1}), glycogen (854 cm^{-1}), and tyrosine (858 cm^{-1}) were regarded as potential biomarkers for the nonexistence of breast malignancy (Bilal *et al.*, 2017). A study by Vargas-obieta *et al.*, (2016) showed mixed biochemical contributions from control and breast cancer serum patients primarily at 622 cm^{-1} (phenylalanine), 642 cm^{-1} (tyrosine), 695 cm^{-1} (polysaccharides), 714 cm^{-1} (polysaccharides), 742 cm^{-1} (lipids), 754 cm^{-1} (tryptophan), 875 (tryptophan), and $1,083 \text{ cm}^{-1}$ (phospholipids), 1002 cm^{-1} (phenylalanine), $1,155 \text{ cm}^{-1}$ (β carotene), 1328 cm^{-1} (tryptophan), and 1556 cm^{-1} (tryptophan), though control serum spectrum depicted higher amounts of carotenoids components (1002 cm^{-1} , 1155 cm^{-1} , 1167 cm^{-1} , 1523 cm^{-1}). It can be concluded that biochemical changes due to proteins were predominant during breast malignancy. Moreover, a decrement in carotenoid levels with cancer is a sign of increased carotenoid degeneration. A limitation with these reported blood-based Raman studies for breast malignancy detection is lack of a chemical / morphological model aimed at quantifying the observed biomarker alterations.

2.3 Leukemia

Leukemia is predominantly a blood and bone marrow cancer, classified (according to rate of growth and cell type) as either acute lymphocytic (ALL), chronic lymphocytic (CLL), acute myeloid (AML), or chronic myeloid (CML) (American Cancer Society, 2017). By 2013, leukemia had risen to tenth place in terms of cancer incidence and ninth place in terms of cancer deaths, with 1 / 127 men versus 1 / 203 women developing leukemia between birth and age of up to 79 years (Naghavi, 2015). It accounts for 29% of all childhood cancers in children aged 0 to 14, but most cases are diagnosed in adults aged 20 and up, with CLL (37%) and AML being the most common forms (31%) (American Cancer Society, 2017).

Previous reviews have extensively demonstrated applicability of blood-based FTIR and Raman spectroscopy in leukemia studies (Wang *et al.*, 2014; Kendall *et al.*, 2009), suggesting blood is a key biofluid for leukemia diagnostics. For instance, a previous serum-based leukemia study observed increased levels of carotenoid and protein in control samples when compared to diseased patients (MartinEspinoza *et al.*, 2008). In addition, there was absence 853 cm^{-1} (protein) band in leukemic spectrum when compared to control spectrum. Application of PCA and LDA showed machine learning techniques' strength in differentiating spectra of the normal and diseased groups. Elsewhere, spectroscopic work based on leukemic cell lines (lymphoma, lymphoid, myeloid leukemia cells) and incorporating multivariate statistical techniques of PCA and LDA observed major spectral differences occurring in $4000 - 400 \text{ cm}^{-1}$ region with heightened nucleic

acid contents observed in myeloid cell lines, while lymphoma cell line had decreased levels of amide proteins (Babrah *et al.*, 2007). Also, another blood-based leukemia study on normal (control) and patients suffering from chronic lymphocytic leukemia (CLL) employing cluster multivariate analysis algorithms was reported (Erukhimovitch *et al.*, 2006), where the results showed biochemical alterations of deoxyribose, phospholipids, and DNA significantly reduced in all normal patients. Moreover, the cluster analysis at these specific regions demonstrated clear discrimination between diseased and healthy patients.

However, to the best of author's knowledge, saliva-based leukemia studies employing Raman spectroscopy as a vibrational technique have not been reported, although saliva-based leukemic studies in enzyme-linked immunosorbent assay methods are well established. For instance, DNA genomic study on leukemia performed on saliva and urine samples showed saliva could be useful for leukemia diagnosis (Rasi *et al.*, 2011). Recently, it was shown that leukemic signatures can be detected from salivary RNA, results that agreed to those obtained from the bone marrow (Chen *et al.*, 2014). Indeed, a comprehensive review of known elevated circulating miRNAs associated with chronic lymphocytic leukemia, based on several genome and transcriptome studies has suggested the salivary μ RNAs could be associated with presence of leukemia (Allegra *et al.*, 2012). Moreover, the salivary 92 miRNA was shown to increase especially in acute myeloid leukemia patients (Allegra *et al.*, 2012). It is believed the differentiation of nucleic acids components play a key role during cancer proliferation (Rasi *et al.*, 2011). Notably, other studies have shown salivary amylase as a potential marker for presence and progression of leukemia. For instance, when comparing leukemic patients to controls, leukemic patients displayed higher levels of amylase and total proteins in their saliva (Ashok *et al.*, 2010). This partly agreed with other work (Singh *et al.*, 2014), that reported elevated saliva amylase levels in leukemic patients when compared to control ones. It is thought that increased amylase levels in body fluids may be an indication of inflammation of the body organs for example, the pancreas, due to underlying medical conditions (Ashok *et al.*, 2010; Singh *et al.*, 2014).

2.4 Machine learning-fluid biomarkers detection approach for cancer diagnosis

Based on above findings, it is evidently clear that there lies great possibility of exploring the potential blood and salivary biomarkers for diagnosis of cancers. However, saliva has a range of drawbacks some of which are shared with other sources of biomarkers like blood, and others that are more unique to salivary gland physiology (Pernot *et al.*, 2014). For instance, composition

of saliva (biomarkers) and flow can be influenced by many factors, for example, smoking, oral diseases and stimulation methods (Shirtcliff *et al.*, 2000). In addition, brushing teeth may cause blood to leak into the saliva, resulting in an increase in molecule concentration (Kivlighan, *et al.*, 2005). This demonstrates importance of homogenizing time and saliva sampling conditions (Pernot *et al.*, 2014), a vital consideration that was taken into account during saliva collection in this study.

A key challenge in bio-spectroscopy is that the analytes in blood and saliva exist in low concentrations (Pfaffe *et al.*, 2011; Christodoulides *et al.*, 2005). Previous studies have shown biochemical components in whole blood and saliva in a human body exist in various ranges of concentrations. For instance, concentrations of protein, lipids, DNA, RNA and saccharides components in blood range in following limits; protein: 6-8 g / dl, lipids: 35-135 mg / dl, DNA: 14-17 mg / l, RNA: 144-166 mg / l, and saccharides: 80-120 mg / dl. For saliva, protein: 0.72-2.45 mg / ml, lipids: 0.9-1.3 mg / dl, DNA: 1 – 100 ng / μ l, RNA: 4,912 – 15,473 ng / μ l, and saccharides: 0.005-0.01 mg / ml (Saroch *et al.*, 2012; Hughes *et al.*, 2019; Poehls *et al.*, 2018; Brozoski *et al.*, 2017; Jurysta *et al.*, 2009; Panchbhai, 2012; Id *et al.*, 2020; Mcmenamy *et al.*, 1957; Gahan, 2010; Leeman *et al.*, 2018). It should however be noted that plasma and saliva biomarkers may not be directly correlated to each other because salivary biomarkers can be produced locally, for example in gum, or the salivary glands (Pernot *et al.*, 2014). As suggested in previous studies, the molecule's concentration in saliva is usually much lower than in blood (Pernot *et al.*, 2014). For instance, saliva contains plasma steroids in the range of 1% to 10%, a factor attributed to binding of carrier proteins (Fraumeni, 2011). It should however be noted that other biofluids in direct contact with diseased tissue, such as sputum, pancreatic juice, ascitic, cerebro-spinal, and urine, bile fluids, are of great importance as media to detect biomarkers that are secreted or shed locally (Baker *et al.*, 2016). It would be expected that the biomarkers should be found in these fluids in higher concentrations than in the blood. Furthermore, since local biofluids have a less complex molecular structure than blood, their detection can be conveniently done. Despite these limitations, blood and salivary biomarkers can be used as a complement to conventional screening methods, aided by inclusion of robust machine learning techniques for accurate diagnostics.

With regard to Raman spectroscopy, a key issue is the identification and isolation of subtle biochemical differences in the biofluids (especially at the onset of disease progression) due to their low concentration against elevated background and fluorescence noise. To address this

problem, several techniques have been developed, including the Surface-Enhanced Raman Scattering (SERS), Resonance Raman Scattering (RRS), Tip-Enhanced Raman scattering (TERS), and Coherent Anti-Stokes Raman Scattering (CARS). Specific details regarding these techniques can be found elsewhere (Wachsmann-hogiu *et al.*, 2009; Israelsen *et al.*, 2017). Among these techniques, SERS can easily detect Raman scattering from single molecules, making it suitable for label-free spectroscopy (Wachsmann-hogiu *et al.*, 2009). SERS is also beneficial because of Raman scattering's molecular specificity and high sensitivity in probing subtle molecular profile changes (Avram *et al.*, 2020). As a result, SERS has become the method of choice for leukemia and breast cancer research.

As compared to traditional Raman spectroscopy, SERS on breast cancer cells showed chemical constituents in the cell nucleus and cytoplasm, such as DNA, RNA, and the amino acids tyrosine and phenylalanine, could be detected with improved sensitivity (González-Solís *et al.*, 2013). Furthermore, the observed SERS spectra needed less laser exposure time (≈ 5 seconds) than Raman spectra, which took 40 to 60 seconds to obtain a spectrum with well-defined peaks. Recently, salivary proteins have been discovered to be potentially useful for noninvasive and label-free breast cancer detection (Feng *et al.*, 2015). However, although the first seven factor latent variables (LVs) accounted for ≈ 92 percent of total variance, only two discriminant functions (LV1 and LV2) specifically distinguished levels of malignancy, and the remaining latent variables were discarded. Vargas-Obieta *et al.*, (2016) demonstrated that SERS and PCA-LDA could be used to differentiate between breast cancer and control samples with high sensitivity and specificity using serum samples. Furthermore, biomolecules such as phenylalanine, tryptophan, carotene, tyrosine, glutathione, amide I, and amide III were identified at low concentrations thanks to the strongly enhanced Raman bands in the $600\text{--}1800\text{ cm}^{-1}$ region. However, as opposed to notion that the first principal components accounting for largest variances could have been useful for spectral discrimination, only the seventh (PC7), eighth (PC8), and tenth (PC10) principal components allowed the best discrimination between breast cancer and control serum samples in the region $600\text{--}1800\text{ cm}^{-1}$ region, suggesting that higher-order principal components maybe potentially useful for spectra discrimination. SERS of serum was also used to detect bio-molecular variations at various stages of breast cancer, with PCA-LDA of SERS spectra being found to differentiate healthy from breast cancer patients with sensitivity of 92 percent and specificity of 85 percent, as well as subjects with breast cancer at various stages with diagnostic efficiency of sensitivity and specificity of 80% (Cervo *et al.*, 2015). The distinction between the SERS spectra of the control

and diseased groups was clearly defined by the scores discrimination due to the first two discriminant functions. The prominent metabolites differences, especially uric acid and hypoxanthine at 721 cm^{-1} , 1093 cm^{-1} , 1324 cm^{-1} , and 1444 cm^{-1} , could be attributed to this separation. As a result, it's unclear if further analysis on the weak/subtle metabolites detected by high-order PCs could have beneficial in spectra discrimination.

Similarly, the SERS technique has been a critical diagnostic method in the study of leukemia onset and progression. For example, SERS-based quantification of leukemia cells spiked in control cells revealed that the key differences between monocyte lysates from three healthy donors and lysates from the leukemia cell line THP-1 were primarily due to the protein to nucleic acid ratio within each cell type (Hassoun *et al.*, 2018). Also, a previous study demonstrated hematologic malignancy and chronic lymphocytic leukemia could be detected using SERS gold nanoparticles (Nguyen *et al.*, 2010). Aside from the SERS signal remaining high after incubation in Eosin, Hemotoxylin, and Giemsa stains, the gold nanoparticles were found to exhibit a persistent strong Raman enhancement for samples after one month. These findings suggests SERS possessed significant advantage over fluorescence in probing chronic lymphocytic leukemia (CLL) cells. In a separate analysis, SERS of DNA derived from an acute myeloid leukemia (AML) cell line showed a lower intensity of 5-methylcytosine (1005 cm^{-1}) than standard DNA (Moisoiu *et al.*, 2019). The findings showed that cancer DNA's methylation pattern affects DNA adsorption geometry, resulting in higher adenine SERS intensities for cancer DNA. The PCA-LDA employing the first two principal components (PC1, PC2) that accounted for 68% variance yielded an overall accuracy of 82.2%, and sensitivity of 75%. It is observed that the remaining PCs that accounted for significant amount of variance (32%) were discarded, suggesting that further analysis on discarded PCs could have been potentially useful in revealing other significant subtle bio-chemical components for spectra discrimination. The human myeloid leukemia cells (HL60, K562) were recently differentiated from normal human bone marrow mononuclear cells using a combination of electroporation-based SERS technique, PCA-LDA, and PLS diagnostic algorithms (BMC) (Yu *et al.*, 2017). PCA-LDA had a sensitivity and specificity of 98.3 percent and 98.3 percent, respectively, in distinguishing leukemia cell SERS spectra from normal cell SERS spectra, while the PLS approach had a diagnostic accuracy of 96.7 percent in predicting unidentified subjects. However, PCA-LDA concentrated on the first three PCs, which accounted for 83.5 percent of the variance: PC 1 (52.876 %), PC 2 (28.976 %), and PC 4 (1.650 %). As a result, it's

unclear if the remaining PCs (16.5 %) may have revealed loading vectors (biochemical assignments) which may have been useful for cell diagnostic classification.

The utility of higher-order principal components is one possible solution for mining potentially important subtle biochemical alterations for disease diagnostics. This technique has not been explored in the conventional Raman and SERS studies of breast cancer and leukemia. Principal component analysis (PCA) is a commonly used multivariate analysis technique for eliminating redundancy in original datasets, and has chiefly featured in diagnosis of breast cancer (Feng *et al.*, 2015; Nargis *et al.*, 2019; Bilal *et al.*, 2017; Vargas-Obieta *et al.*, 2016; Cervo *et al.*, 2015) and leukemia (MartinEspinoza *et al.*, 2008; Babrah *et al.*, 2007; Moisoiu *et al.*, 2019; YU *et al.*, 2017). The aim of PCA is to reduce dimensionality while allowing for as much variance in the original data set as possible (Martinez *et al.*, 2005). However, only the few principal components that explain much variance are retained while the remaining PCs (high-order PCs) are discarded (Jolliffe, 2002). As a result, in the presence of much higher variance signals, the low-variance signals (in the case of the present study weak Raman bands corresponding to trace and ultra-trace disease biomarkers) may become lost in the noise of higher principal components (Pelletier, 2003). According to Jolliffe (2002), analyzing intermediate and high-order principal components can be useful in obtaining additional knowledge about between-group variation. This is in line with findings by Pelletier (2003) who suggested that analyzing the omitted higher principal components (intermediate- and higher-order principal components) may be a useful tool for detecting analyte information.

A major limitation with PCA is the requirement that the data should lie on linear subspace (Luo *et al.*, 2008). Thus, although PCA make the loading vectors uncorrelated to each other, the components may not be statistically independent. Therefore, spectrum components may interact with one another, making it difficult to determine their precise concentrations of the constituent biochemical components. In this regard, Independent Component Analysis (ICA) has been suggested as a better alternative to PCA, owing to its ability of optimizing independence conditions to give more meaningful components (Masood *et al.*, 2006). Furthermore, ICA works with non-Gaussian data, producing physically observable concentration profiles and spectral components. With such capability, the low variant signals (subtle occurrences and multivariate alterations) determined with the help of high-order principal components can be further explored using ICA to ensure their mutual independence during quantification. To the best of author's knowledge, this

has not been previously done in breast and leukemia studies. However, it should be noted that the performance is dependent on the correct choice of ICA model dimensionality (Taleb *et al.*, 2006). There are many versions of ICA which include fast independent component analysis (FASTICA), joint approximate diagonalization of eigenmatrices (JADE), kernel ICA (KICA), Infomax ICA and Mean-field ICA (MF-ICA) (Wang *et al.*, 2008; Boiret *et al.*, 2014). Amongst the ICA versions, the FASTICA algorithm is a suitable choice for blood and saliva spectral data analysis owing to its ability of estimating only certain desired ICs, rather than solving the entire mixing matrix (Wang *et al.*, 2008).

Previous works on blood and saliva-based breast cancer and leukemia studies show PCA was used alongside LDA (Nargis *et al.*, 2019; Bilal *et al.*, 2017; Vargas-Obieta *et al.*, 2016; MartinEspinoza *et al.*, 2008; Babrah *et al.*, 2007). Being a low dimensional classifier, LDA require already dimensionally reduced data, and works efficiently well with linearly separable feature space (Wang *et al.*, 2014). In event of high-nonlinear datasets, SVM, and ANN would be suitable techniques for pattern recognition and nonlinear regressions; a task that has hardly been explored in aforementioned breast cancer and leukemia works. It should however be noted that evolutionary genetic classification algorithms are known to suffer numerical instability especially when no proper initialization is provided. SVM has been the most popular kernel - based machine learning algorithm for both regression and classification purposes (Ratle *et al.*, 2010; Archibald *et al.* 2007; Switonski *et al.*, 2010). It deals with two-class problems where maximum performance is required. Key challenges with SVM include i) it requires lengthy training time, ii) its outputs represent selections rather than future probabilities, iii) formulation of classes greater than two is usually problematic, and iv) predictions must be certain / positive (Wang *et al.*, 2014; Bishop *et al.*, 2006).

The optimal choice of kernel parameters has been suggested as a potential method of achieving better performance with ANN and SVM (Wang *et al.*, 2014). For SVM, a dataset with a low training samples to input dimensionality ratio and high dimensional data provide the best generalization efficiency (Belousov *et al.*, 2002). However, the selection of an optimal kernel function for a given task or clear guidelines for using specific kernels is still of research interest. The efficiency of an SVM classifier is largely determined by the kernel it employs, taking into account size and speed constraints in both training and testing. For saliva and blood-based breast cancer and leukemia studies, a potential method of achieving greater SVM classifier's performance, and which has not been experimented in aforementioned studies could be testing a set of cost values. The cost value with the lowest cross-validation classification error can then be

chosen as the best for further data analysis. For ANN, the training of neural network classifiers necessitates a lot of computation and is plagued by issues like convergence to local minima and outlier vulnerability (Tu, 1996). The high nonlinearity associated with neural network methods may result in trapping in local minima and thus limiting their discrimination power (Tang *et al.*, 2009). Furthermore, since neural networks are extremely sensitive to the Hughes phenomenon, they are ineffective when dealing with a large number of spectral bands (> 41) (Camps-valls *et al.*, 2005). To get around this problem, one possible solution is to use error feedback from the training samples to adjust the weights, bringing the network prediction of the correct outputs for the training samples closer to the true values. In particular, experimentation can be performed by adjusting the number of neurons per layer, the learning rate, alpha values and the number of iterations.

There is currently a scarcity of studies using novelized machine learning methods, such as low variance principal components, to investigate useful subtle markers for breast cancer and leukemia diagnosis. Further, previous studies have not realized quantification of the potentially useful low variance signals. Moreover, the diagnostic usefulness of determined low variance signals are not definitely clear, particularly where maximization of generalization performance is required in a two-class problem and in a multiclass problem where data nonlinearity is more likely to occur, for example, in ANN classification problem.

Chapter 3 Principles of Quantitative Raman spectroscopy

3.1 Basics of Raman spectroscopy

Upon interaction of electromagnetic radiation with a material, the possible processes include absorption, transmission, reflection and scattering (Fotakis *et al.*, 2007). Absorption process happens when the transitions between bound states are in resonance with the incident photons. The bound transition states could be vibrational, electronic, and rotational. Based on electromagnetic spectrum, electronic transitions take place in the visible, ultraviolet, and near-infrared regions, while rotational and vibrational take place in the far-infrared and infrared regions, respectively (Fotakis *et al.*, 2007). Light transmission shows low (weak) to no absorption. Scattering of electromagnetic radiation can be categorized into elastic / Rayleigh scattering or inelastic / Raman scattering. The energy of dispersed radiation in Rayleigh scattering has the same wavelength (and frequency) as the incident radiation. The higher and low scattered frequencies are categorized into anti-Stokes and Stokes Raman, respectively, thanks to Sir Chandrasekhara Venkata Raman (1888-1970) who discovered the Raman effect in 1928 (Chandra *et al.*, 2015).

Raman spectroscopy utilizes scattering phenomenon to excite targeted molecules into the vibrationally excited state (Fotakis *et al.*, 2007; Matthäus *et al.*, 2008), where the anti-Stokes and Stokes scattering can be visualized in terms of photon energy gain or loss as depicted in Figure 3.1. The molecule is excited from its ground state into a virtual excited state by incident monochromatic radiation at frequency ν_0 (and energy = $h\nu_0$). In the case of Rayleigh scattering, the molecule relaxes back to its ground state and releases a photon of equal energy. In Stokes Raman scattering, the molecule relaxes to an excited vibrational level of the ground electronic state, releasing a photon whose energy (frequency: $\nu_S = \nu_0 - \nu_v$) is lower by the corresponding vibrational continuum. Since $\nu_v < \nu_0$, there is an increment in size of wavelength (λ). It can also be seen (Figure 3.1) that the molecule transitions from an excited vibrational level of the ground electronic state to a virtual state in anti-Stokes Raman scattering, then relaxes to the ground level, releasing a photon of energy increased by one vibrational quantum (frequency: $\nu_{AS} = \nu_0 + \nu_v$). Since $\nu_v > \nu_0$, there is a decrement in size of wavelength (λ) shortens.

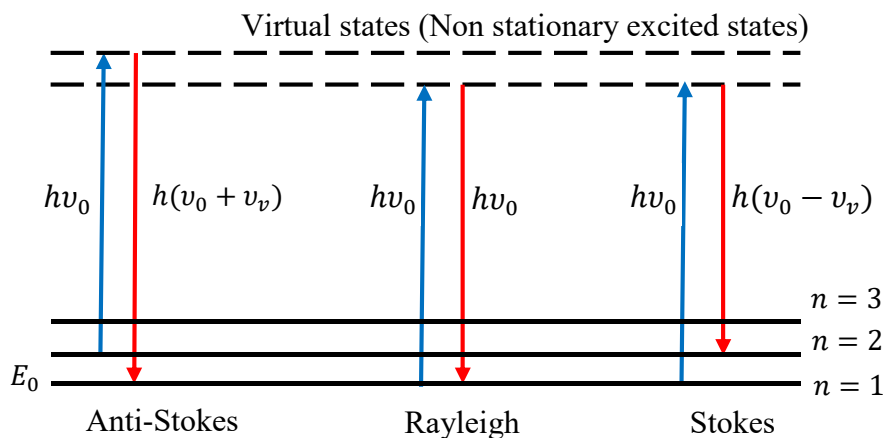


Figure 3.1 Diagram of energy levels demonstrating processes of anti-Stokes, Rayleigh, and anti-Stokes Raman scattering (Fotakis *et al.*, 2007).

Raman spectrum is obtained upon spectrally resolving the scattered radiation (Fotakis *et al.*, 2007). This is a graph of dispersed / scattered light intensity as a function of the difference in frequency between incident and scattered radiation ($\Delta\nu$) / Raman shift (in cm^{-1} units). The observed Raman bands reveal vibrational modes that explain unique fingerprints of the sample materials under investigation (Matthäus *et al.*, 2008). Furthermore, since molecular vibrations are strongly influenced by conformation of the molecule and its chemical environment (Schie, 2013), spectral analysis may assist in identifying active molecular bands and assessing the impact of the microenvironment on studied samples, for example, biological cells.

In Raman spectroscopy, a large percentage of incident photons ($> 99.9\%$) undergo elastic Rayleigh scattering. Therefore, only a small percentage of incident light ($\leq 0.001\%$) undergo inelastic scattering (with frequencies $\nu_o \pm \nu_m$), which can be considered useful for molecular characterization. For this reason, a major drawback with spontaneous Raman scattering is its weak signals and which are mostly dominated by broadband fluorescence emissions. For instance, though both fluorescence and Raman spectroscopy are based on vibronic effects, Raman scattering is at least 6 orders of magnitude weaker than fluorescence (Matthäus *et al.*, 2008). Nevertheless, Raman spectroscopy (including microspectroscopy) possess key features that make it attractive to scientists (Fotakis *et al.*, 2007): First, the technique is extremely precise due to the unique fingerprint aspect of the Raman spectrum. Furthermore, Raman microscopes have excellent spatial resolution, which aids in the study of small features, for example, in sub-cellular analysis. Raman spectroscopy is also nondestructive, and can be performed *in situ* thus minimizing sampling time.

Other positive attributes include: availability of advanced fiber-optic Raman probes for remote sample analysis, availability of Raman databases for identification of unknown sample components (e.g. in soil and mineral analysis), and easy versatility usage nature of Raman set up systems. For instance, Raman microscopes are easy to use and are easily coupled to ones choice of excitation sources, spectrographs and CCD (charge-coupled devices)

3.2 Theory of Raman spectroscopy

The phenomena behind Raman spectroscopy are better understood upon studying the interaction of the photons with the molecule utilizing the Time Independent Schrödinger equation. Time-independent Schrödinger equation for free particles with energy E_0 is given by (Schwabl, 2007; Abdelrahman *et al.*, 2014):

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi_0}{\partial x^2} = E_0 \psi_0 \quad (3.1)$$

It is observed in equation (3.1) that the potential vanishes since the particles are free. The solution of this equation is

$$\psi_0 = e^{iax} \quad (3.2)$$

which can be rewritten as $\psi_0 = A_0 \cos ax$ (3.3)

Substituting equation (3.3) in equation (3.2), we obtain

$$\frac{\partial \psi_0}{\partial x} = -a A_0 \sin ax \quad (3.4)$$

and

$$\frac{\partial^2 \psi_0}{\partial x^2} = -a^2 A_0 \cos ax = -a^2 \psi_0 \quad (3.5)$$

Therefore, equation (3.3) becomes

$$-\frac{\hbar^2 a^2}{2m} \psi_0 = E_0 \psi_0 \quad (3.6)$$

And because

$$E_0 = \frac{\hbar^2 k^2}{2m} \quad (3.7)$$

then

$$-\frac{\hbar^2 a^2}{2m} \psi_0 = \frac{\hbar^2 k^2}{2m} \psi_0 \quad (3.8)$$

Now, $a = k$. Therefore,

$$\psi_o = A_o \cos k_o x \quad (3.9)$$

Substituting the Eigen function ψ_P of the photon in Schrödinger equation (3.3) the equation becomes

$$-\frac{\hbar^2}{2m} \frac{\partial^2 \psi_P}{\partial x^2} = E_P \psi_P \quad (3.10)$$

Solving equation (3.10), we obtain wave equation in equation (3.11)

$$\psi_P = A_P \cos k_P x \quad (3.11)$$

Upon interaction of electrons of the molecule and photons, we obtain a wave equation ψ_T as follows

$$\psi_T = \psi_o \psi_P \quad (3.12)$$

$$\text{which can be rewritten as, } \psi_T = A_o A_P \cos k_o x \cos k_P x \quad (3.13)$$

Using trigonometric relations and substituting $A_o A_P$ by A_T , equation (3.13) can be rewritten as

$$\psi_T = \frac{A^T}{2} [\cos(k_o + k_p) x + \cos(k_o - k_p) x] \quad (3.14)$$

In equation (3.14), $(k_o + k_p)$ represents the photons' absorption by the electron while $(k_o - k_p)$ represents the photon's emission by the electron.

Including equation (3.13) into equation (3.14), we obtain

$$\psi_T = \frac{A^T}{2} [\cos(k_o + k_p) x + \cos(k_o - k_p) x] + A_P \cos k_P x \quad (3.16)$$

It can be observed (equation 3.16) that there are three wavelengths $\lambda_o + \lambda_p$, $\lambda_o - \lambda_p$, and λ_p , which represent anti-Stokes Raman, Stokes-Raman and Rayleigh scattering, respectively.

Most molecules are contained in the ground state at room temperature. According to Boltzmann's theorem, the Stokes Raman lines are far more intense than the anti-Stokes Raman lines in thermal equilibrium (Larkin, 2011). It's worth noting that the polarization of the scattered beam might not be identical to that of the incident beam (Smith *et al.*, 2005). As noted (Figure 3.2), Raman spectrometers contain analyzer component for analyzing the polarization of the scattered beam. Besides, it has polarizer component that i) ensures radiation is plane polarized and ii) determines the angle of the plane of the incident radiation. The analyzer allows polarized beam to pass through only in one plane, first by allowing transmission of scattered radiation in the plane

of the incident radiation (parallel scattering) and then allow beam whose polarization direction has been changed at 90° by the molecule M (perpendicular scattering).

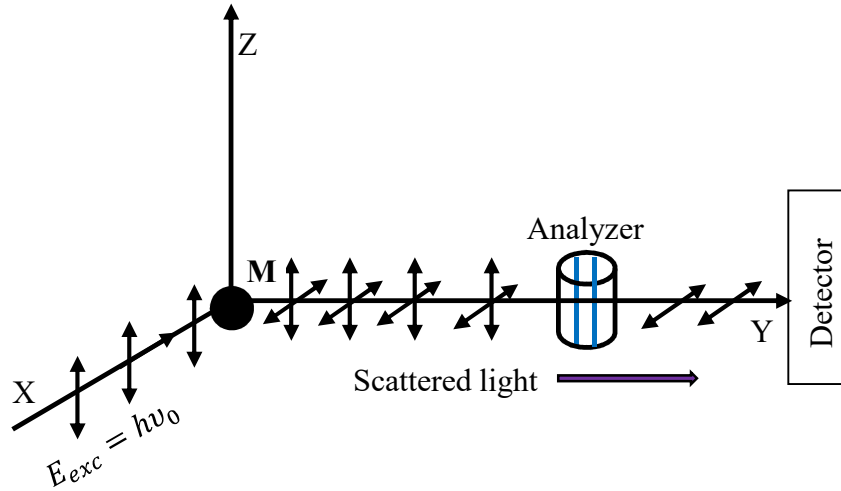


Figure 3.2 The schematic structure of a Raman instrument.

The intensity of Raman scattered radiation (I_{ij}) is given by (Larkin, 2011):

$$I_{ij} = K v_s^4 P_{ij}^2 N_i I_0 \propto v^4 I_0 N \left(\frac{\partial \alpha}{\partial Q} \right)^2 \quad (3.17)$$

where K = factors of instrumentation

v_s = scattered photon frequency,

P_{ij} = the transition probability,

$N_i = N$ = concentration of scattering molecules,

I_0 = intensity of incident radiation (laser intensity),

v = frequency of incident radiation (exciting laser),

α = polarizability of the molecules,

Q = vibrational amplitude.

Equation (3.17) implies that i) quantification of molecular signatures is possible since Raman signal is concentration dependent, ii) Raman intensity of the various molecular bands increases by using shorter wavelength excitation or increasing the laser flux power density, and iii) only molecular vibrations which cause a change in polarizability are Raman active as governed by:

$$\left(\frac{\partial\alpha}{\partial Q}\right)_0 \neq 0. \quad (3.18)$$

Another observation concerns the parallel connection between Infrared (IR) absorption and Raman scattering (Schrader, 1995; Lewis *et al.*, 2001). IR absorption is a one-photon effect that occurs when the frequency of IR radiation and the vibrational frequency of a specific normal mode of vibration are in direct resonance. As a result, the dipole moment of the molecule changes in relation to its vibrational motion. In this case, the IR photon interacts with the molecule, the photon vanishes, and the molecule's vibrational energy is raised by the photon's energy at the vibrational resonance frequency. Raman scattering, on the other hand, is a two-photon reaction involving a polarizability change of the molecule in relation to its vibrational motion. Electrons and nuclei are forced to travel in opposite directions when a molecule is exposed to an electric field. As polarizability interacts with incoming radiation, it induces a dipole moment in the molecule that is proportional to the electric field intensity and molecular polarizability α . The radiation emitted by this induced dipole moment contains the observed Raman scattering.

3.3 Raman spectrometric instrumentation

A Raman spectrometer simply consists of excitation source (in this study, a laser source), optical filters, charge-coupled device (CCD), and a spectrograph / spectrometer (Figure 3.3). Briefly, the laser light illuminates the sample. The optical filters consists of excitation and emission filters. Narrow band-pass filters are used to block laser noise in the excitation filters. The emission filter is made up of an edge long-pass filter that suppresses Rayleigh light while allowing scattered Raman signals to pass through to the Raman spectrograph, imaging spectrograph, and CCD camera. The illumination pinhole effectively creates a single point source, which the objective refocuses onto the sample to analyze each point on the sample. The confocal pinhole functions as a spatial filter, allowing only in-focus light to pass through while effectively eliminating out-of-focus light from the specimen. When the image point on the detector has the same focus as the illumination light spot on the target plane, the object and images are said to be confocal (i.e., the object point and the image point lie on the optically conjugated planes). Several laser mirrors guide the laser beam to the sample, which is then centered onto the sample using a microscope objective. The backscattered light is captured by the objective lens and guided to the spectrograph's entrance slit. The notch filter (or edge long-pass filter) reduces the transmission of intense Rayleigh lines

through the spectrograph. The scattered radiation is then reported on a CCD array detector using a diffraction grating spectrograph.

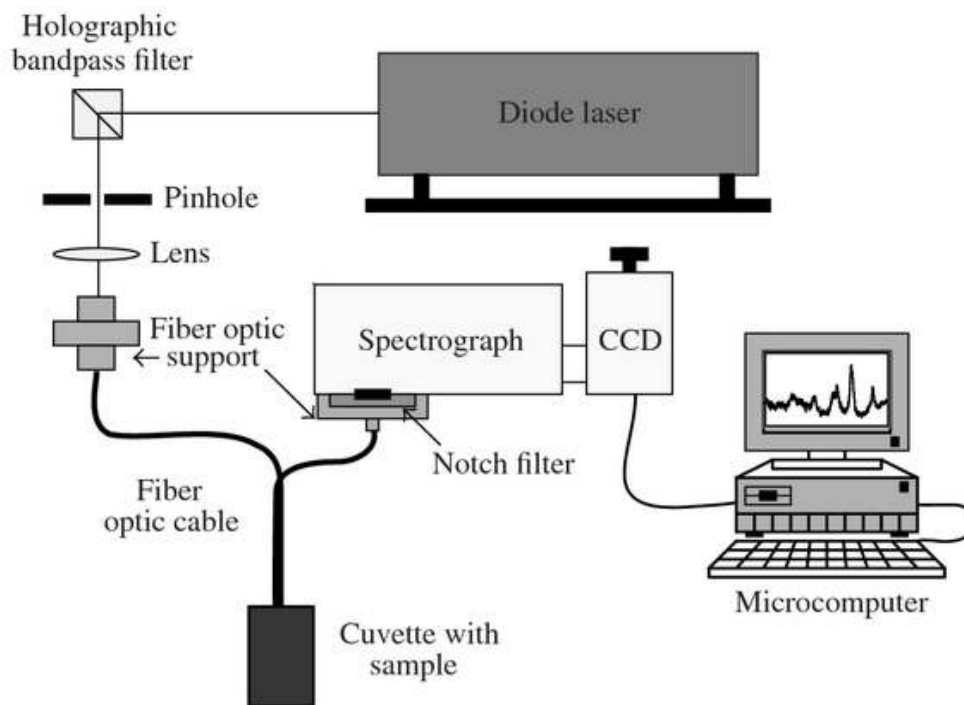


Figure 3.3 Schematic illustration of a typical Raman spectrometric set up.

The choice of particular laser is one of the most important consideration in any spectroscopic study (Schie, 2013; Byrne *et al.*, 2015). From equation (3.17), Raman lines intensity (I) depends on laser frequency's fourth power (ν) where ($I \propto \nu^4$). As a consequence, at shorter wavelengths, scattering efficiency is higher, resulting in faster integration times (Byrne *et al.*, 2015). However, while shorter wavelength laser sources e.g., 532 nm, improve signal efficiency (and thus detector performance), they also increase sample photodegradation and autofluorescence (Schie, 2013). Moreover, utility of higher wavelength sources (in μm scale) pose challenge of lower quantum efficiency of detectors and significantly reduced cross section (Schie, 2013). The most common compromise therefore include utilizing near-infrared (NIR) excitation sources such as 785 nm, 830 nm and 1064 nm. Such excitation sources have good sample (for example, tissues) penetration depths thereby exciting large volumes of sample (Wang *et al.*, 2014). Moreover, excitation at longer wavelengths (785 and 1064 nm) is frequently used as optimal for measuring of fresh tissues due to relatively low background (Synytsya *et al.*, 2014). In general, 785 nm is the

most preferred excitation laser wavelength in bio spectroscopy studies because it provides a balance of performance with less excitation efficiency but also lower fluorescence (Synytsya *et al.*, 2014).

3.4 Calibration regression for quantitative Raman spectral analysis

The most widely used regression methods for quantification of multivariate spectral data are partial least-squares (PLS), ANN, principal component regression (PCR), multivariate linear regression (MLR) variants, ridge regression (RR), continuum regression (CR) and support vector regression (SVR) (Geladi *et al.*, 2016; Salem *et al.*, 2014; Awad *et al.*, 2015). The ANN and SVR regression techniques work very well in nonlinear situations and with large heterogeneous data sets (Geladi *et al.*, 2016; Awad *et al.*, 2015). It is based on the perceptron – the so considered basic building block of ANN. Perceptron is the name initially given to a binary classifier. However, perceptron can be viewed as a function which takes certain inputs and produces a linear equation which is nothing but a straight line (Rezaeianzadeh, 2014). Generally, sigmoid function or other similar classification algorithms, for example, linear kernel, polynomial kernel, and Gaussian radial basis function (RBF) are used as activation functions. The process involves feeding input to a neuron in the next layer to produce an output using an activation function (Rezaeianzadeh, 2014). This process is called as ‘feed forward’. After producing the output, error (or loss) is calculated and a correction is sent back in the network. This process is called as ‘back propagation’ (Rezaeianzadeh, 2014). The back propagation process aims at achieving the smallest training error as a function of the added neurons to the intermediate layer (Okonda *et al.*, 2017). The back-propagation ANNs are the most useful for calibration purposes (Geladi *et al.*, 2016). However, in comparison to the traditional linear regression techniques e.g., partial least-squares (PLS), ANN regression often give less stable and less robust results (Geladi *et al.*, 2016).

Support Vector Regression; a method that uses the same principle as the SVMs, is also applicable for quantifying large nonlinear heterogeneous data sets. The objective of support vector regression is to basically consider the points that are within the decision boundary line. The best fit line is the hyperplane that has a maximum number of points (Salem *et al.*, 2014; Awad *et al.*, 2015). In other words, the straight line that is required to fit the data is referred to as hyperplane. The data points on either side of the hyperplane that are closest to the hyperplane are called support vectors (Salem *et al.*, 2014). Unlike other regression models that try to minimize the error between the real and predicted value, the support vector regression tries to fit the best line within a threshold

value Suárez *et al.*, 2011). The threshold value is the distance between the hyperplane and boundary line (Salem *et al.*, 2014). The fit time complexity of support vector regression is more than quadratic with the number of samples which makes it hard to scale to datasets with more than a couple of 10000 samples. Support vector regression technique has many advantages which include (Awad *et al.*, 2015): being robust to outliers, easy update of decision model, excellent generalization capability, high prediction accuracy, and easy implementation. However, they have shortcomings which include (Awad *et al.*, 2015): (i) They are not suitable for large datasets, (ii) They underperform in cases where the number of features for each data point exceeds the number of training data samples, and (iii) The decision model does not perform very well when the data set has more noise i.e. target classes are overlapping.

PLS regression is a commonly preferred technique in biospectroscopy. The PLS regression model can be thought of as a combination of principal component regression (PCR) and multivariate linear regression (MLR) (Allegrini *et al.*, 2014). The PLS regression is a common technique for quantitative Raman spectroscopy. The aim of quantitative Raman spectral analysis is to deduce the composition from the Raman spectrum of a sample. Based on working principal of PCR and MLR (Allegrini *et al.*, 2014), PLS can be interpreted to be a factorial analysis that simultaneously take both spectral and biochemical data into account (Weinmann *et al.*, 1998). PLS has an advantage over PCR in that it needs fewer components in the model, making it easier to identify outliers and groupings in datasets (Geladi *et al.*, 2016). PLS regression does not entail a priori knowledge of the spectra of all the components in a complex mixture, unlike other spectrum analysis approaches. Moreover, the success of PLS regression is pegged on establishing a suitable efficient model during the calibration phase, whose optimal number of factors to be used are determined by validating model (Weinmann *et al.*, 1998). Further details regarding conceptual principles of PLS regression can be found elsewhere (Geladi *et al.*, 2016; Valderrama *et al.*, 2007).

For spectroscopy works, multivariate analysis of PLS regression involves solving the equation $X = C(c.S) + E$ where X = original spectra (measured spectrum), $c.S$ = spectra of calibration samples, C = concentrations matrix to be projected, and E = residuals (Stone *et al.*, 2007). The equation provides the best fit of basis spectra found within the measured spectrum. It further assumes residual is held to a minimum so that the key components of measured spectra are the chosen spectral components (Stone *et al.*, 2007). Careful consideration should be taken into account when choosing the number of models to include in the model. Selecting too few

components results in a poor model, while selecting too many components results in a model that is susceptible to noise (Geladi *et al.*, 2016).

The fit of the model can be evaluated by scatter plotting the values for the test set against the measured values for model validation. A diagonal scatter plot should be the product of a good model. Alternatively, the linearity can be used to estimate the model qualitatively by comparing the distribution of residuals to the reference values (Frost, 2016). In addition, the accuracy of the model can be evaluated by determining the coefficient of multiple determination during cross-validation (R^2_{val}) and the root mean squared error of prediction ($RMSEP$), according to the following equation (Frost, 2016; Sichangi *et al.*, 2018).

$$RMSEP = \sqrt{\frac{\sum_{i=1}^n (y_{pred,i} - y_i)^2}{n}} \quad (3.19)$$

where $y_{pred,i}$ refers to predicted concentration; y_i is the known concentration for sample i and n is the number of samples.

3.5 Multivariate machine learning techniques for Raman spectral analysis

To better understand biochemical differences between samples categories, machine learning techniques are performed with aim of revealing similarities and differences in signature profiles embedded in overlapping spectra.

3.5.1 Principal component analysis (PCA)

The principal component analysis (PCA) is a popular data dimensionality reduction technique aimed at extracting the smallest number of principal components that represent most important information in the original multivariate data. Although there are many variants of PCA algorithms, the singular value decomposition principal component analysis (SVD-PCA) is often preferred in reducing redundancy of the spectral information owing to its matrix factorization algorithm flexibility (Martinez *et al.*, 2005; Trauth, 2015). The SVD-PCA is a chemometric method that seeks to obtain a lower rank approximation to matrix \mathbf{X} (Martinez *et al.*, 2005),

$$\mathbf{X} = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T \quad (3.20)$$

where \mathbf{U}_k is a $n \times k$ matrix containing the first k columns of \mathbf{U} , \mathbf{V}_k is the $p \times k$ matrix whose columns are the first k columns of \mathbf{V} , and \mathbf{D}_k is a $k \times k$ diagonal matrix whose diagonal elements are the k largest singular values of \mathbf{X} . The approximation is usually the best one in least squares sense; to

provide principal components (PCs) that contain the main amount of variance (potential information) pertaining to the data. The first principal components (loads) that describe the greatest variance from the mean is often utilized during subsequent analysis (Crow *et al.*, 2005).

3.5.2 Linear discriminant analysis (LDA)

The goal of the LDA technique is to project the original data matrix onto a lower dimensional space (Tharwat *et al.*, 2017). Being a low dimensional classifier, LDA require already dimensionally reduced data, for instance by PCA, and fits well with feature space that is linearly separable (Li *et al.*, 2012). The LDA technique is developed to transform the features into a lower dimensional space, which maximizes the ratio of the between-class variance to the within-class variance, thereby guaranteeing maximum class separability (Li *et al.*, 2012; Tharwat *et al.*, 2017). LDA follows three steps to achieve dimensional reduction of datasets (Varmuza *et al.*, 2008; Tharwat *et al.*, 2017): (i) calculation of the between-class variance or between-class matrix, (ii) calculation of the distance between the mean and the samples of each class, i.e., the within-class variance or within-class matrix, and (iii) construction of the lower dimensional space which maximizes the between-class variance and minimizes the within-class variance.

3.5.3 Independent component analysis (ICA)

ICA is a blind source separation algorithm that seeks to find a nonlinear representation of non- Gaussian data such that, independent components (ICs) highly correlated to the spectral profiles of components in the mixtures are extracted (Wang *et al.*, 2008; Trauth, 2015). ICA aims at estimating k independent vectors (independent components) from a noise-free model (Boiret *et al.*, 2014);

$$X = A.S \quad (3.21)$$

where X is a $(n \times m)$ matrix, S is a $(k \times m)$ matrix of k independent independent components, and A is a $(n \times k)$ matrix of coefficients of X . A computed matrix U constituted by the independent components would therefore be given by

$$U = W.X = W.(A.S) = S \quad (3.22)$$

There are many versions of ICA such as fast independent independent component analysis (FASTICA), joint approximate diagonalization of eigenmatrices (JADE), kernel ICA (KICA), Infomax ICA and Mean-field ICA (MF-ICA) (Wang *et al.*, 2008; Boiret *et al.*, 2014). The fast

independent component analysis (FASTICA) algorithm is often preferred for spectral analysis due to its ability of estimating only certain desired ICs, rather than solving the entire mixing matrix (Wang *et al.*, 2008).

3.5.4 Multidimensional scaling (MDS)

Multidimensional scaling is a technique for the analysis of similarity or dissimilarity data on a set of objects (Leeuw, 2005). MDS attempts to model such data as distances among points in a geometric space (Martinez *et al.*, 2005; Leeuw, 2005). The main reason for doing this is that one wants a graphical display of the structure of the data, one that is much easier to understand than an array of numbers and, moreover, one that displays the essential information in the data, smoothing out noise (Leeuw, 2005). The most widely assumed metric in MDS is the Euclidean, in which the distance between two points j and k is defined as (George *et al.*, 2020):

$$d_{jk} = \left[\sum_{r=1}^R (X_{jr} - X_{kr})^2 \right]^{1/2} \quad (3.23)$$

where X_{jr} and X_{kr} are the r th coordinates of points j and k , respectively, in an R -dimensional spatial representation. It should be noted that two-way multidimensional scaling use either the Euclidean metric or the Minkowski ρ (or L_ρ) metric, which defines distances as (George *et al.*, 2020; Weinberg, 1991):

$$d_{jk} = \left[\sum_{r=1}^R (X_{jr} - X_{kr})^\rho \right]^{1/\rho} \quad (\rho \geq 1) \quad (3.24)$$

Equation (3.24) includes Euclidean distance as a special case in which $\rho = 2$.

3.5.5 Support vector machine (SVM)

Support vector machine is a supervised machine learning algorithm that can be used for classification and regression problems. The working principle of SVM is based on the fitting function (Han *et al.*, 2017):

$$f(x) = \sum_{i=1}^n \alpha_i y_i K(x, x_i) + \beta_o \quad (3.25)$$

where $K(x, x_i)$ is a kernel function; x_i is the training sample eigenvector; and x is the recognizing sample eigenvector (Singla *et al.*, 2011). The parameter α_i is restricted to $0 \leq \alpha_i \leq C$ and can be

estimated by maximizing a Lagrangian. C is the cost parameter that determines the amount of regularization, that is, the parameter C (cost function) determines the classification error term (Bouzalmat *et al.*, 2014). There are various SVM functions e.g., the linear and radial basis functions (RBF) kernel functions defined as (Han *et al.*, 2017; Singla *et al.*, 2011):

$$K(x, x_i) = x \cdot x_i \quad (3.26)$$

$$K(x, x_i) = \exp(-\gamma \|x - x_i\|^2) \quad (3.27)$$

In spectroscopy, spectral data analysis looks for the best RBF function parameter γ , as well as the best regularization parameter, C , for the efficient optimization process. The optimal γ and C values are sorted in such a way that the effect is sufficient to create a decision surface without misclassifying the training set, hence reducing over-prediction. The γ and C are obtained from the grid search for the highest cross-validation accuracy.

3.5.6 Back propagation neural network (BPNN)

The backpropagation network belongs to a class of artificial neural networks (ANN). ANNs are biologically inspired computational networks which learn to approximate a unidirectional mapping from an n -dimensional input space R^n to an m -dimensional output space R^m (where n represents the number of input variables, m represents the number of output variables) (Marini *et al.*, 2008). The implementation of backpropagation neural network (BPNN) is based on the fitting function (Marini *et al.*, 200; Wythoff, 1993):

$$y = f\left(\sum_{i=1}^n w_i x_i + \theta\right) \quad (3.28)$$

where y represents the neuron output, i.e., the value of the nonlinear function f (the neuron itself) corresponding to the inputs x_i ; x_i represents the input weights; w_i represents the node bias (offset); and θ represents the number of node synapses. The performance is optimized by minimizing the mean squared error (MSE) (equation 3.29) by adjusting the network weights.

$$MSE = \frac{1}{n} \sum_{i=1}^n \left[\sum_{j=1}^m (y_{ij} - o_{ij})^2 \right] \quad (3.29)$$

where y_{ij} represents the correct pattern outputs for a pattern i ; o_{ij} represents the network estimates for pattern i ; m represents the number of output nodes; and n represents the training patterns counts.

Chapter 4 Materials and Methods

4.1 System Configuration and Optimization

This section outlines the system configuration and optimization procedures that were undertaken during spectral measurements.

4.1.1 STR Raman spectrometer system

The STR series Raman spectrometer was used to record Raman spectra. This spectrometer is available at the Department of Physics, University of Nairobi. Briefly, this setup (Figure 4.1) comprises the 785 nm and 532 nm excitation lasers. The excitation laser beam traverses through 3 meter length, multimodal 10 μm core diameter optical fiber connected to the Raman optics assemblage. The Raman light travels through the 3m length, 50 μm core diameter fiber from Raman optics to the spectrometer. The Raman optics assemblage is coupled with confocal Raman optical microscope system (Olympus BX51, Olympus Corporation, Tokyo). The assemblage houses a combination of neutral density (ND) filters (in various orders of percentage transmissions), band pass filters and long pass filters (all from Semrock Corporation) interfaced to shutter controller, motorized stage controller (MAC 6000 system) and motorized XYZ stage - BP-3''X2''(all from Ludl Electronics Products, Ltd).

Raman signals are detected by combination of 0.3 mm imaging triple grating spectrograph (Princeton Instruments, Acton SpectraPro2300) and a -75°C Pentium cooled Halogen charge-coupled device (CCD). The LL01-532 and LL01-785 neutral density filters are used to reject unwanted lines from 532 nm and 785 nm excitation sources, respectively. Rayleigh scatterings are removed by LP03-532RU and BLP01-785R long- pass edge filters respectively (all from Semrock Corporation). The window-based STR Raman version 1.9.3 software is used to set and manage the experimental parameters.

4.1.1.1 STR Raman spectrometer system power loss characteristics

As described in Section 3.2, the choice of optimum excitation wavelength for biological studies must balance between minimizing induced fluorescence signals and the fourth power dependency (ν^4) of laser frequency(ν) (Larkin, 2011). The best compromise in many research works has been utility of near-infrared (NIR) excitation sources such as 785 nm, 830 nm and

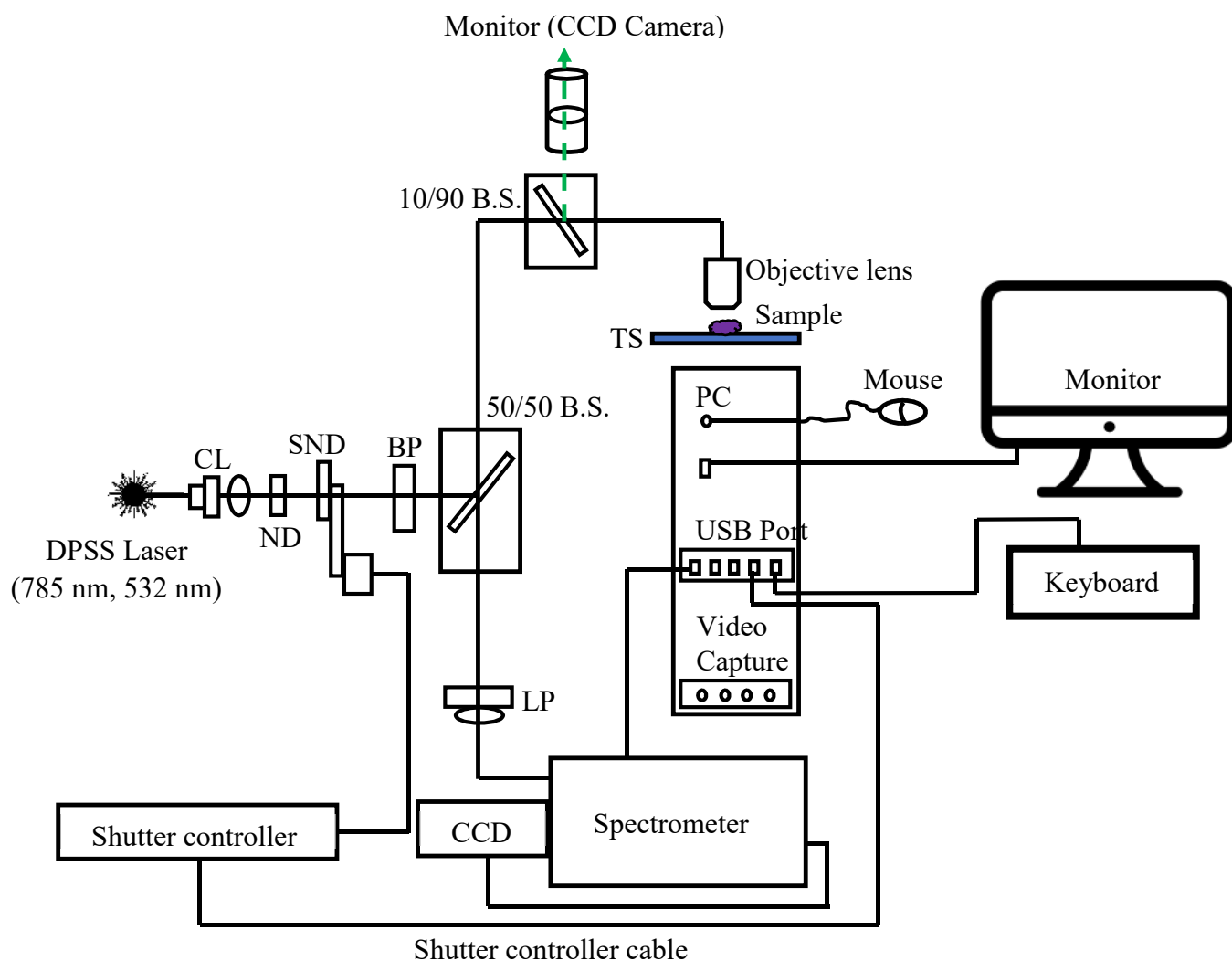


Figure 4.1 Schematic diagram of a customized STR Raman spectrometer system.

KEY: DPSS - diode-pumped solid-state; CL - collimator lens; ND - neutral density filter (1, 5, 10, 25, 50, 100 %); SND - shutter neutral density filter (0.1%, x2); BP - Band pass filter; 50 / 50 B.S. - beam splitter (50 % reflection, 50 % transmission); LPF - long pass filter; 10 / 90, B.S. - Beam splitter (10 % reflection, 90 % transmission); CCD - charge-coupled device; TS - translational stage (XYZ stage); PC - personal computer; USB -universal serial bus.

1064 nm. Therefore, this work was based on 785 nm excitation laser due to following advantages (Zhao *et al.*, 2014): First, the fluorescence background emanating from tissues is relatively lower at 785 nm laser excitation than at shorter excitation wavelengths, for example, the 532 nm wavelength. Secondly, the wavelength at 785 nm excitation penetrates tissues more deeply than shorter wavelengths. Moreover, the tissue's Raman quantum yield is higher at this excitation wavelength than at longer excitation wavelengths.

To get an understanding of energy transfers (and losses) through system components up to the sample on the stage, the 785 nm laser powers were measured after several components i.e., collimator lens (CL), neutral density filters (ND), shutter neutral density filters (SND), band pass filter (BP), flat mirror, 50 / 50 BS (beam splitter; 50 % reflection, 50 % transmission) and the x80 objective lens. This was done using a calibrated Newport photodiode (detector 818-SL) fitted with a neutral density filter (OD3) coupled to a Newport optical power meter (model 840) (power density: $0.1 \text{ mWm}^{-2} - 2000 \text{ Wm}^{-2}$). The respective measured power losses are summarized in Table 4.1.

Table 4.1 The power losses in the STR Raman system's main components

| System components | Measured Power (mW) | Power loss (dB) |
|---|---------------------|-----------------|
| Diode N-IR (785 nm) laser | 100 | --- |
| Collimator lens (CL) | 64.1 | 1.9314 |
| Neutral density filter (ND; 100%) | 64.0 | 0.00678 |
| Shutter neutral density filter (SND) | 31.2 | 3.1202 |
| Flat mirror (reflective) | 24.9 | 0.9795 |
| 50 / 50 beam splitter | 13.58 | 0.2633 |
| x80 ULWD objective (0.80) Spot size: 10.85 μm | 9.8 | 0.1416 |
| x80 LWD objective (0.50) Spot size: 13.32 μm | 11.2 | 0.0836 |

We observed that the largest loss of laser light energy occurred between the laser head and the collimator lens (CL), neutral density filter (ND), and shutter neutral density filter (SND), as shown in Table 4.1. This loss of laser light energy can be attributed to energy losses through optical medium via mechanisms of reflection, absorption, and scattering (Fotakis *et al.*, 2007). After

repeated experiments based on x80 objective, it was observed that the maximum measured power on the sample surface ranged between 9.72 ± 0.02 mW to 11.56 ± 0.32 mW. Consequently, all the subsequent Raman measurements regularly took into account the laser power on sample surface during all actual measurements on biofluid samples.

4.1.1.2 Stability of laser output characteristics

The warm-up and power output stability characteristics of the 785 nm laser were tested over a two-hour period after the laser was turned on. The laser intensity was measured using a calibrated Newport photodiode (detector 818-SL) fitted with a neutral density filter (OD3) coupled to a Newport optical power meter (model 840) (power density: $0.1 \text{ mWm}^{-2} - 2000 \text{ Wm}^{-2}$), every 15 seconds. Similarly, ambient temperature (thermometer: MOD BA-888 Oregon SCIENTIFIC) was measured every 15 seconds. Respective plots of measured laser output and ambient temperature versus time is shown in Figure 4.2.

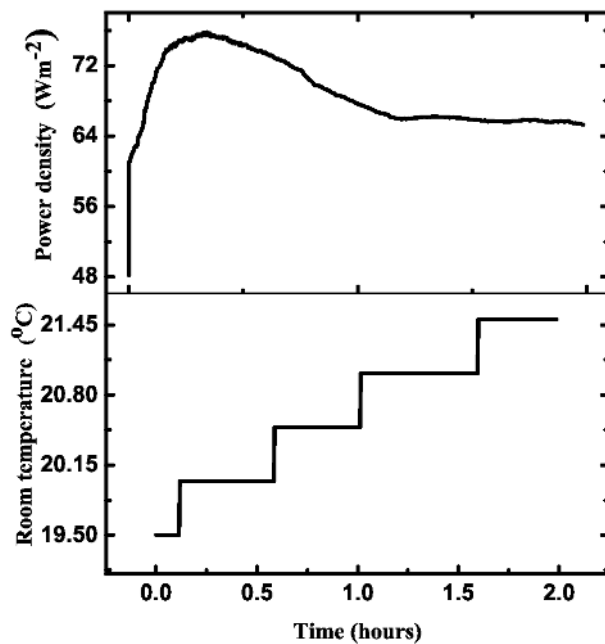


Figure 4.2 Plot of laser output power at 785 nm (in watts / square meter) and ambient temperature (°C) versus time (hours).

It was observed that the laser power (and power density) increased drastically in the first 15 minutes from switching on the laser. This can be attributed to the initial coherent emission when a laser current (I_{LD}) above a certain threshold flows through it. Previous studies have shown that

laser diodes do not have a constant optical output power, particularly in the first 1 hour after being switched on (Semleit *et al.*, 1997; Zhou, 2015). In this study, the laser power density was found to gradually decrease from the 20th up to the 75th minute, where it stabilized over a period of the next two to five hours. This trend was consistently observed after repeating similar measurements in the next three days. Moreover, there was a gradual rise in ambient temperature, which attained approximate levels of 22.29 ± 0.41 °C at the point of laser power stability. The increment of ambient temperature could be attributed to the heat dissipation from the laser and STR series Raman device. However, it should be noted that variations in environmental temperatures might lead to a higher root mean square noise levels in the optical output power, hence lower excitation efficiency (Semleit *et al.*, 1997; Zhou, 2015). Based on this finding, the 785 nm laser was always switched on for about 1.5 hours prior to beginning Raman measurements, and the air conditioner was always set at 22 °C, to ensure constant room temperature, hence stable laser output power. Moreover, Raman measurements were performed within a period of five hours.

4.1.1.3 Wavelength stability of laser

The Raman spectrum of crystalline silicon material was measured every minute during the 2 hour warm-up cycle in order to assess the wavelength stability of the 785 nm laser. The air conditioner was initially set at 22° C and laser was switched on for more than 1.5 hours before measurements. Figure 4.3 (a) shows the measured crystalline silicon spectrum versus time (minutes). The silicon peak was found to have a mean wave position of 519.49 ± 0.0075 cm⁻¹ within the 2 hours measurements (Figure 4.3 (b)). Thus, to ensure wavelength stabilities during subsequent Raman measurements, the STR system was regularly recalibrated after every 2 hours.

4.1.1.4 Choice of optimal substrates for biological samples measurements

The suitability of available substrates for biological spectral measurements in our laboratory, which included the ordinary microscopic glass slides, silver-coated glass slides, and calcium fluoride substrates were evaluated. Using 785 nm excitation laser, a total of 5 spectra, exposure time = 50 seconds were measured and averaged for graphical analysis. Fig. 4.4 shows the average spectrums of ordinary glass slides, silver-coated glass slide, and calcium fluoride substrates. It can be observed that the ordinary glass fluoresces highly fluoresced around 1380 cm⁻¹ band, and was therefore found unsuitable for Raman measurements in the 400-1800 cm⁻¹ region. The silver paint coated glass and calcium fluoride substrates had minimal fluorescence,

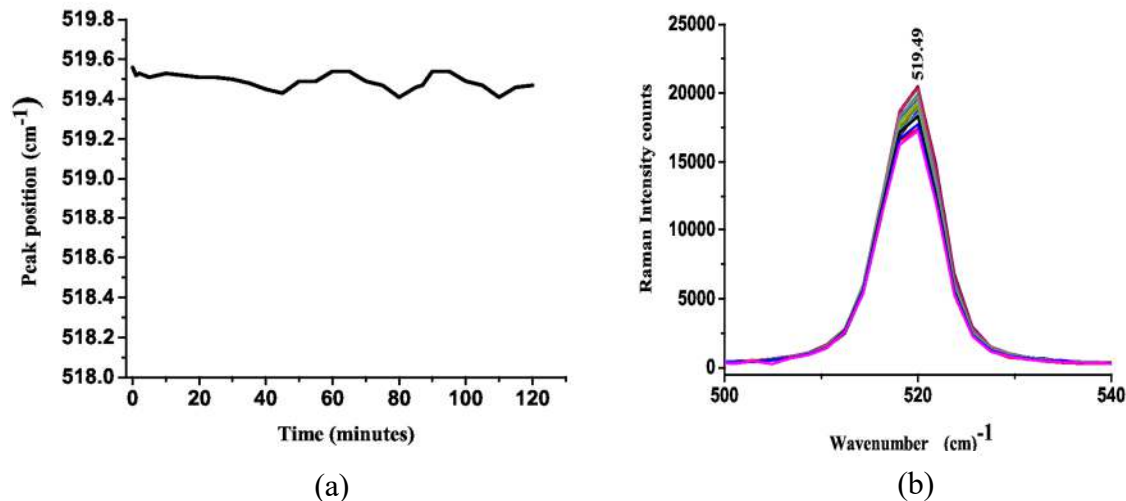


Figure 4.3 (a) The measured crystalline silicon position at 519 - 520 cm⁻¹ against time (in minutes), and (b) mean wavenumber position of crystalline silicon (519.49 ± 0.0075 cm⁻¹). The power on sample surface ≈ 9.8 ± 0.26 Mw, spot size ≈ 10.85 μm.

although calcium fluoride substrate performed better than all the other substrates, which agrees with other findings (Byrne *et al.*, 2015). In this study, it was observed that the silver-coated glass slides could not be used for spectral measurements of saliva, because saliva deposits could not be practically identified on the substrate surface. However, silver-coated glass slides yielded enhanced and clearer Raman spectra from blood fluids. Moreover, the calcium fluoride was observed to be a better substrate owing to its Raman peaks free characteristics in the 400-1800 cm⁻¹ region (Byrne *et al.*, 2015). Therefore, the current study undertook Raman measurements on blood and saliva samples using silver coated glass slides and calcium fluoride substrates, respectively.

4.1.1.5 Effects of ambient light on measurements

The effects of ambient light on spectral measurements were evaluated by measuring the Raman spectrum of calcium fluoride substrate in 150-1900 cm⁻¹ region, under several experimental conditions which included: when fluorescent lights were switched on or off, halogen light intensities set at maximum and minimum values, and desktop screen away and close to the sample area. Prior to taking Raman measurements, the 785 nm laser was switched on for approximately 2 hours, and the air conditioner was constantly set at 22 °C. This ensured laser power stability during

the measurements. To improve signal intensity, the substrates were illuminated with 785 nm laser at exposure times 100 seconds. Figure 4.5 shows the effects of ambient light on spectral curves of calcium fluoride substrate.

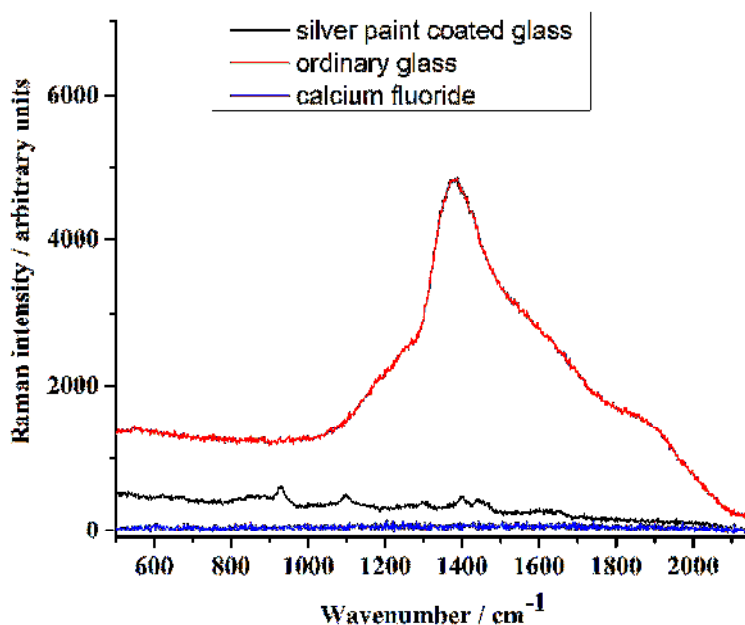


Figure 4.4 Average spectrums of ordinary glass slides, silver coated glass slide, and calcium fluoride substrates; laser power on sample surface 10.1 ± 0.04 mW; spot size ≈ 10.85 μm , exposure time = 50 seconds.

About eight Raman peaks (320, 363, 381, 411, 428, 1266, 1341, 1535 cm^{-1}) were observed, as shown in Figure 4.5. The intense peaks at 381 cm^{-1} , 411 cm^{-1} , 1266 cm^{-1} , 1341 cm^{-1} , and 1535 cm^{-1} points to spectral contributions from fluorescent lamps whereas the (363 cm^{-1} , 428 cm^{-1}) bands explain the spectral contributions from tungsten halogen lamps (Zhao *et al.*, 2014; Desroches *et al.*, 2015). The peak at 320 cm^{-1} is specific to Raman grade calcium fluoride substrate materials (Lewis *et al.*, 2017). Based on this data, the subsequent Raman measurements were limited to 400-1800 cm^{-1} region and taken in complete darkness. Nevertheless, it was impossible to turn off the computer's liquid crystal display (LCD) monitor. However, the LCD monitor was always kept pointing away from the point of measurements, and not too close to where measurements were made. The LCD monitor did not seem to have a noticeable impact on the measured spectra, according to a visual analysis of the measured spectra.

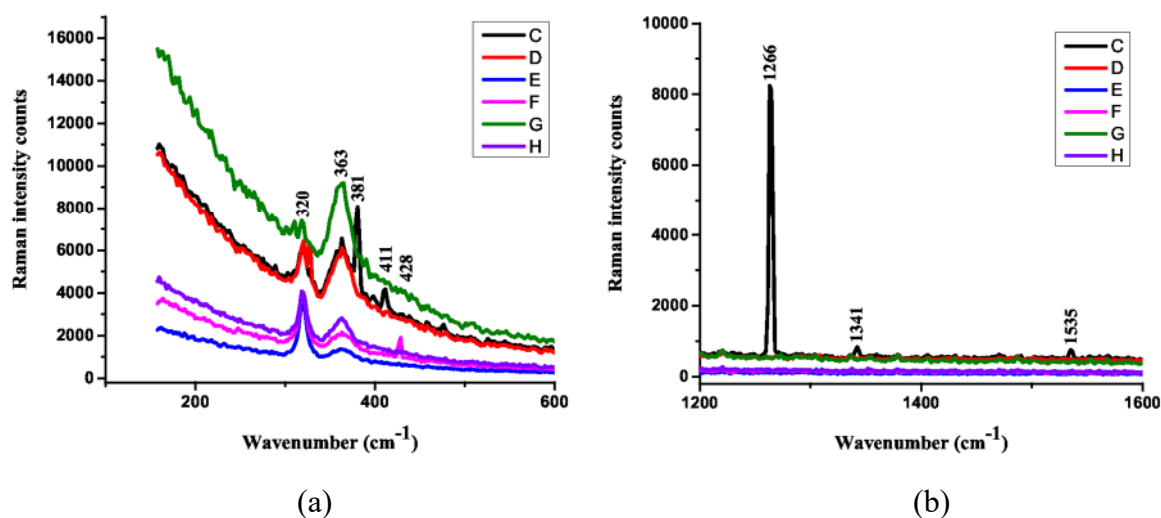


Figure 4.5 Spectrum showing the effects of ambient light on spectral curve of calcium fluoride substrate in (a) 200 – 600 cm^{-1} , and (b) 1200 – 1600 cm^{-1} regions.

KEY: C – ambient light switched on, halogen light and epi-illumination set at maximum intensity / level; D - ambient light switched off but halogen light and and epi-illumination set at maximum intensity / level; E - ambient light switched off, halogen light set at maximum intensity / level and epi - illumination switched off; F - ambient light switched off, halogen light set at minimum value and epi - illumination switched off; G- ambient light switched off, halogen light set at minimum value and epi - illumination set at maximum value; H – ambient light switched off, halogen light set at minimum value, epi illumination switched off and desktop screen light blocked from sample surface.

4.1.1.6 The effect of power density and exposure times on signal-to-noise ratio (*SNR*).

The effects of exposure times and power levels on *SNR* of Raman peaks were evaluated by measuring spectra of blood and saliva samples using an x80 objective at ≈ 12 mW laser power on the sample surface. The resulting averaged spectra of 10 measured spectra of blood and saliva are shown (Figure 4.6 (a), (b)).

The *SNR* values were calculated based on equation (4.1) (McCreery, 2001) for different integration times at chosen Raman bands, namely (1241 cm^{-1} , 1445 cm^{-1}) for blood samples, and (1268 cm^{-1} , 1539 cm^{-1}) for saliva samples. These bands were selected for determination of *SNR* values because they provide biochemical information regarding protein to lipid ratios in biological samples (Movasaghi *et al.*, 2007).

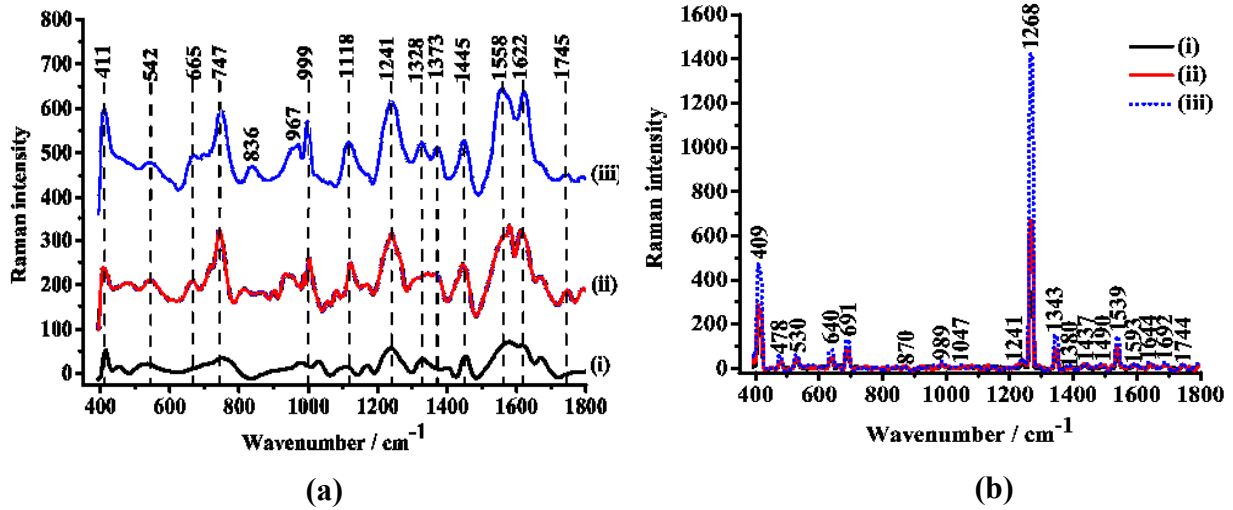


Figure 4.6 Raman spectra of (a) blood, and (b) saliva from a grade 3 breast cancer patients, demonstrating the effect of exposure times on *SNR* of selected bands, in total exposure times of 60 seconds (i), 90 seconds (ii), and 120 seconds (iii).

$$SNR = \frac{\bar{S}}{\sigma_y} \quad (4.1)$$

The \bar{S} and σ_y in equation (4.1) represented the average peak height and the peak height's standard deviation, respectively. The respective *SNR* ratios are summarized in Table 4.2. Evaluation of *SNR* values showed the exposure time of 120 seconds delivered the optimal intensity signal without burning the sample. The sample damage was determined by checking for burning marks on the sample after measurements.

Table 4.2 The *SNR* values of saliva and blood Raman peaks measured at different exposure times

| Wavelength / nm = 785 nm | | | | |
|----------------------------|--|------|--------|------|
| | Blood | | Saliva | |
| Exposure time (seconds) | Signal-to-Noise Ratio at wavenumber shift (in cm^{-1}) | | | |
| | 1241 | 1445 | 1268 | 1539 |
| 60 | 2.3 | 1.7 | 8.9 | 0.8 |
| 90 | 4.2 | 2.6 | 10.7 | 1.1 |
| 120 | 9.8 | 5.6 | 14.3 | 3.5 |

As seen in Figure 4.6 (a), it is evidently clear that blood spectra is dominated by hemoglobin and fibrin which are protein components that constitute building blocks of red blood cells and plasma, respectively (Lbany *et al.*, 2015). For instance, the 999 cm^{-1} , 1373 cm^{-1} , and 1622 cm^{-1} are characteristic peaks of hemoglobin. The 967 cm^{-1} and 1241 cm^{-1} peaks can be attributed to pure fibrin, which is a major component of coagulated blood. The 1241 cm^{-1} band reveals amide III vibrational modes of proteins in form of C–N stretching, N–H bending, and C–C stretching (Rehman *et al.*, 2013). Based on the available literature (Rehman *et al.*, 2013; Sikirzhyski *et al.*, 2010), the biochemical alterations in saliva spectra (Figure 4.6 (b)) can be assigned to cholesterol (418 cm^{-1}), cholesterol esters / cysteine (539 cm^{-1}), cholesterol esters / C–C twisting of proteins (617 cm^{-1}), nucleic acids (655 cm^{-1} , 722 cm^{-1} , 1096 cm^{-1} , 1140 cm^{-1} , 1185 cm^{-1} , 1281 cm^{-1} , 1426 cm^{-1} , 1511 cm^{-1} , 1679 cm^{-1}), amino acids (793 cm^{-1} , 978 cm^{-1} , 1598 cm^{-1}), phosphodiester (824 cm^{-1}), phospholipids / proteins (878 cm^{-1}), proteins / glycogen (937 cm^{-1}), amide III (1281 cm^{-1}), CH_3CH_2 wagging modes in collagen and purine nucleic acid bases (1324 cm^{-1}), lipids / proteins (1382 cm^{-1} , 1549 cm^{-1}), and lipid esters (1744 cm^{-1}).

4.2 Cell culture: *in vitro* application in cancer research

In the last decade, *in vitro* cancer research has relied heavily on animal cell culture. The word "cell culture" refers to the process of growing eukaryotic, prokaryotic, or plant cells under controlled conditions. Therefore, animal cell culture refer to propagation of cells derived from animal cells (Freshney, 2006). The cells may have been directly extracted from tissue and disaggregated by enzymatic means before cultivation, or they may have been derived from a previously identified cell line or cell strain (Freshney, 2006). The cells are normally passaged to provide more space for continued growth by moving them to a new vessel with fresh growth medium.

For cell proliferation to occur, artificial formulations / environments including supplies of essential nutrients (vitamins, carbohydrates, amino acids, and minerals), growth factors, hormones and essential gases (carbon dioxide and oxygen) are used (Masters *et al.*, 2007). Carbohydrates are supplemented in form of glucose. In some instances, it is replaced with galactose to decrease lactic acid build up, because galactose is metabolized at a slower rate. Other carbohydrates sources include amino acids (particularly L-glutamine) and pyruvate. In addition to nutrients, tissue culture media contain bicarbonate that necessitate a 5% carbon dioxide atmosphere to maintain the pH at 7.4 and osmolality of culture system (Masters *et al.*, 2007). The pH is maintained by one or

more of buffering systems; the CO_2 or / and Na_2CO_3 . Most mammalian cell lines grow well at pH 7.2 - 7.4 (Masters *et al.*, 2007). The commonly used media are Eagles Minimum Essential Medium (EMEM) and Dulbecco's Modified Eagle's Medium (DMEM). The EMEM and DMEM may generally be classified as complete culture media (basal media). These are media supplemented with other media additives such as serum, L-glutamine, and antibiotics. When complete media is supplemented with other percentage formulations of fetal bovine serum, antibiotics, fungizone and L-glutamine, a new media hereby referred to as growth media is obtained. Growth media is used for initiating growth of cells in tubes. It is worth noting that L-glutamine is an important amino acid that is necessary for protein synthesis, energy production, and nucleic acid metabolism in virtually all mammalian cells grown in culture (Masters *et al.*, 2007).

Amino acids, proteins, vitamins, carbohydrates, lipids, hormones, growth factors, minerals, and trace elements are all contained in animal serum. Serum acts as a buffer for the culture medium, inhibits proteolytic enzymes, and increases the viscosity of the medium. In addition, serum provide hormonal factors for stimulating cell growth and proliferation, promotes differentiated functions, provides transport proteins, and enhances attachment and spreading factors (Rauch *et al.*, 2009; Brunner *et al.*, 2009). Sera from fetal and calf bovine sources are commonly used to support the growth of cells in culture (Brunner *et al.*, 2009). A cocktail of almost all factors required for cell attachment; termed as Fetal Bovine Serum (FBS), is commercially available for propagation of human and animal cells.

4.2.1 PC3 and PNT1a cells preparation

First, a spectrometric analysis of biochemical changes in a model tissue during cancer initiation and proliferation was carried out. The aim of this preliminary study was to evaluate how intermediate and high-order principal components can be useful at detecting subtle biochemical changes in biological samples' Raman spectra. Due to unavailability of suitable breast and leukemic cell lines, the metastatic androgen insensitive (PC3) and immortalized normal (PNT1a) human prostate cell lines were chosen for a model tissue Raman spectroscopy analysis.

The metastatic androgen insensitive (PC3) and immortalized normal (PNT1a) human prostate cell lines at passages (subculture levels) 3 and 4, respectively, were prepared at Kenya Medical Research Institute (KEMRI), Nairobi. Both cell lines were cultured in Dulbecco's Modified Eagle Medium (DMEM (1x))(Gibco® by Life Technologies™, USA), supplemented with 10 %, by volume, fetal bovine serum (FBS) (ATCC® 30-2020™, USA), 1 % Fungizone

Amphotericin B, 250 µg/ml (Gibco® by Life Technologies™, UK), 1 % L-glutamine 200Mm (100x) (Gibco® by Life Technologies™, USA), 1 % penicillin-streptomycin and 0.1 % Gentamycin solution 50 mg/ml (all from Sigma® Life Science, USA), at 37 °C in a humidified atmosphere containing 5% CO₂. Each cell line was simultaneously proliferated in three batches at seeding densities of 1.5x10⁶, 1.2x10⁶ and 1.0x10⁶ viable cells/ml, whose cells were harvested after 48 hours (stage 1), 72 hours (stage 2), and 96 hours (stage 3), respectively. Further, sterilized calcium fluoride substrates (Crystran, UK) were introduced into T-75 plug seal capped cell culture flasks (Corning® Flask) to allow cell attachments. It should be noted that cell culture flasks have surface activated growth area surfaces treated with specific chemicals e.g., poly-d-lysine or collagen 1, for optimal adhesion and best proliferation of adherent cells. Nevertheless, it was expected that proliferating cells could naturally attach on calcium fluoride substrates. Therefore, no coating was done on substrates to enhance PC3 and PNT1a cells proliferation and adhesion.

Cell harvesting was done by first removing the substrates (containing attached monolayer cells). The substrates were briefly washed in the Hanks' balanced salt solution, fixated in acetone ≥ 99.5 % A.C.S. reagent (Sigma® Life Science, USA) for approximately 8 minutes, rinsed with double distilled Millipore water, then allowed to dry overnight in the biosafety chamber. The attached monolayer cells were later examined by visual observation using the Raman optical microscope system. The remaining cells in T-75 flasks were detached by the usual trypsinization procedure, then centrifuged twice at 1200 revolutions per minute (r. p. m.) for 5 minutes. The resultant cells were twice washed in a 1.5 ml Phosphate-buffered saline solution (Sigma-Aldrich®, USA) and centrifuged at 1200 r. p. m. for 5 min after each wash. After removing remaining supernatants, cells were vortexed again, then suspended in 1.5 ml Phosphate-buffered saline solution and stored at -80° C.

4.3 Blood and saliva samples collection

A group of 23 healthy volunteers / controls (age 34-56 years) and 20 malignant patients (age 41-65 years) all-female participated in breast cancer study while a group of 18 healthy volunteers / controls (age 20-45 years) and 9 malignant patients (age 24-72 years) both females and males participated in the leukemia study. All healthy (control) and diseased volunteers were from the Kenyatta National Hospital (KNH), Kenya. To take part in the study, all of the years old participants signed a written consent form. The Kenyatta National Hospital – University of Nairobi (KNH – UoN) Ethics and Research Committee granted permission for the study (ERC certificate

number: P112 / 03 / 2018). The specimen collection were performed by Dr. Daniel K. Ojuka, Prof. Jessie N. Githanga, and Peninah Kabethi, all from the Kenyatta National Hospital.

For clarity, the recruited healthy volunteer / control(s) patients were the consenting patients admitted at KNH as suspected breast and leukemic cancerous cases but diagnosed to be non-malignant cases (for breast cancer) and non-leukemic cases, respectively. In contrast, the recruited consenting breast cancer and leukemia patients (diseased cases) were patients admitted for suspected breast or leukemia cancers illness and diagnosed to have malignant tumors (for breast cancer) and leukemia, respectively.

About 2 ml volumes of peripheral venous blood and unstimulated saliva were collected from participants during morning hours (9-11 am). Saliva specimens containing traces of blood were immediately discarded and recollected. Every specimen was put into a sterile test tube. All tubes were disinfected before collection of fluid. Ethylenediaminetetraacetic acid was used as anticoagulation agent for blood samples. Blood and saliva test tubes were placed in walk-in freezers and transported to the Kenya Medical Research Institute (KEMRI) laboratory at a temperature of $\leq 4^{\circ}\text{C}$. The period of transportation was no longer than two hours which guaranteed biological properties of samples were preserved (Chiappin *et al.*, 2007).

Different from prostatic cells whose staging was categorized according to the periods of cell harvesting, i.e., 48 hours (stage 1), 72 hours (stage 2), and 96 hours (stage 3), blood and saliva samples were categorized according to the status (healthy, diseased) of the consenting participants. First, the pathology reports of the consenting participants were reviewed. For diseased patients, the staging and histological grading of the breast tumor and leukemia was done according to the World Health Organization guidelines. For breast cancer, tumor grading was accomplished using the Nottingham modification of the Scarff-Bloom-Richardson grading system (Elston *et al.*, 1991). In addition, tumor staging was reported using the TNM system adopted by UICC and the American Joint Committee on Cancer (Sobin *et al.*, 2009). For leukemia, malignancy levels were staged and graded using the French-American-British (FAB) and the morphologic or cytochemical differentiation (AML) classification systems (Hong *et al.*, 2017; Krause, 2000). The classification was done by accredited pathologists experienced in breast and / or leukemia pathology. In this study, malignancies were classified into grade 1 cancer (early malignancy), grade 2 cancer (medium malignancy), and grade 3 cancer (late malignancy). In line with this classification, and for identification during spectroscopic analysis, samples collected from patients diagnosed with grade 1 cancer, grade 2 cancer, and grade 3 cancer were categorized as ‘grade 1 cancer’, ‘grade

2 cancer', and 'grade 3 cancer', respectively. The samples collected from healthy volunteers were categorized as 'controls'. The details regarding samples for breast malignancy and leukemia are provided in Appendices I and II, respectively. The biofluid samples were refrigerated at -20° C, and spectral measurements were performed within a period of 48 hours. For the sample collection, the study applied the criteria outlined in section 4.3.1.

4.3.1 Inclusion and exclusion criteria

4.3.1.1 Breast cancer

Inclusion criteria

Cancer patients: Age ≥ 18 at the time of signing the informed consent form, with confirmed case of breast cancer, or a candidate to mastectomy, and not on chemotherapy.

Controls: Age ≥ 18 at the time of signing the informed consent form, without breast diseases history or detectable abnormalities by self-examination.

Exclusion criteria

Cancer patients: with benign tumors or already treated, and diagnosed with any other malignancies within 5 years.

Controls: with breast diseases history and detectable abnormalities by self-examination (enlarged supraclavicular lymph nodes, painful lumps, thickening and dimpling of the breast skin, change in size of affected breast, change in size of affected breast, redness, swelling and increased warmth (in inflammatory breast), crusting, ulcers or scaling on the nipple, bloody discharge from nipple, signs from metastasis (respiratory (pleural effusion, consolidation), abdomen (jaundice, hepatomegaly, ascites), neuromuscular (headache, seizure, papilloedema)).

4.3.1.2 Leukemia

Inclusion criteria

Patients: Age ≥ 18 at the time of signing the informed consent form, with any confirmed case of leukemia, and not on chemotherapy.

Volunteers: Without suspected case of leukemia after undergoing complete blood picture test.

Exclusion criteria

Patients: Already treated of leukemia cancer, or treated with any other form of malignancy within five years.

Volunteers: With any form of leukemia cancer, or any other form of malignancy.

4.4 Sample preparation

4.4.1 Whole blood and saliva biofluids samples

The frozen whole blood and saliva aliquots were retrieved from the freezing chambers and thawed at approximately 15 °C for 15 minutes. To extract oral mucous epithelial cells and food waste, thawed saliva samples were centrifuged for 5 minutes at 7000 revolutions per minute (Li *et al.*, 2012). No further processing was performed on blood samples. About 5 µl blood and saliva drops were deposited onto freshly sterile prepared silver coated glass and CaF₂ substrate, respectively. For the next three hours, the sample drops were allowed to dry fully in the biosafety chamber at 15 °C. After that, the investigations began at Spectroscopy and Imaging Laboratory, Department of Physics.

4.4.2 Whole blood and saliva simulates

In this study, blood tissue equivalent (base matrix) was prepared using 22% high purity triolein in water (by weight), xanthan gum (Sigma® Life Science, USA), and 10 µm nylon particles of density $\approx 1.032 \text{ g/ cm}^3$ (Orgasol™) at Kenya Medical Research Institute, Nairobi (Ng *et al.*, 2019). As defined in section 4.4.3, the mixture was stirred to ensure a homogeneous solution and spiked with prepared concentrations of various biochemical components (see section 4.4.3) in the range of 1-500 ppm. This range was chosen based on the typically known concentration ranges of biochemical components i.e., proteins, lipids, DNA, RNA, and saccharides in whole blood and saliva in a human body as detailed in Table 4.3. For saliva simulate, Biochemazone™ artificial saliva (pH = 6.8) was used. The saliva was then spiked with prepared concentrations of various biochemical components in the range of 1-500 ppm, detailed in section 4.4.3.

4.4.3 Calibration set design for biochemical components formulation

Previous studies have showed that Raman spectroscopy may provide a methodology for noninvasive detection of diseases by quantifying the biochemicals which are present in normal and diseased tissues, such as proteins, lipids, and nucleic acids, with accurate information for classifying and grading malignancy (Stone *et al.*, 2007; Byrne *et al.*, 2020; Jr *et al.*, 2014). This involves modeling spectra of specific chosen biochemical components that represent biochemical

changes of proteins, lipids, nucleic acids and saccharides e.g., albumen, actin, collgane type 1, elastin, leucine, DNA, glycogen, blood (hemoglobin), phosphatidylcholine (phospholipids), and beta-carotene, in order to estimate the “Raman concentration,” i.e., the relative concentration based on the Raman scattering of each molecule in the tissue spectrum (Stone *et al.*, 2007; Byrne *et al.*, 2020; Jr *et al.*, 2014).

For quantification purposes, pure basic biochemical components which included ribonucleic acid (RNA) extract from whole blood sample ($\geq 99\%$), bovine serum albumin ($\geq 98\%$), glycogen type IX from bovine liver ($\geq 85\%$), glycerol trioleate derived from glycerol ($\geq 99\%$), triolein, L-glutamic acid potassium salt monohydrate ($\geq 99\%$), and glycine ($\geq 98.5\%$) were employed in this study. The chosen pure components represented nucleic acids, amino acids, proteins, polysaccharides, and lipid compounds in cellular constituents. In this study, the albumin was chosen to represent proteins whereas glutamate and glycine represented amino acids. It should be noted that plasma protein is chiefly made of albumin component (Ong *et al.*, 2012). Glycine is a biosynthetic precursor to porphyrins used in red blood cells, whereas polysaccharides in cells are mainly in form of glycogen (Ong *et al.*, 2012). Besides, glycine and glutamine are precursors of nucleotides (Berg *et al.*, 2012). To represent membranous and non-membranous lipids, triolein and glycerol were chosen, respectively (Ong *et al.*, 2012). The extraction process of RNA component was performed by QIAGEN[®] extraction method (GmbH, 2010), at KEMRI, Nairobi. All the other components were purchased from Gibco by Life Technologies[™], USA and Sigma[®] Life Science, USA.

To make fully-dissolved condensed stocks, the pure basic biochemical components were weighed and diluted to specific volumes with distilled water. Based on expected typical ranges of proteins, lipids, DNA, RNA and saccharides in whole blood and saliva of human body (Table 4.3), whole blood and saliva simulates were spiked with prepared concentrations of the various biochemical components, in the range of 1-500 ppm (Table 4.4). The mixtures were stirred for few minutes to ensure a homogeneous solution. Twenty five samples (mixed concentrations) were prepared and frozen at -20°C for spectroscopic measurements.

Table 4.3 Typical concentration ranges of biochemical components in whole blood and saliva in a human body (Saroch *et al.*, 2012; Hughes *et al.*, 2019; Poehls *et al.*, 2018; Brozoski *et al.*, 2017; Jurysta *et al.*, 2009; Panchbhai, 2012; Id *et al.*, 2020; Mcmenamy *et al.*, 1957; Gahan, 2010; Leeman *et al.*, 2018)

| | Blood | Saliva |
|-------------|--|---|
| Components | Concentrations (g / dl, mg / dl, mg / l, mg / ml, ng / μ l, ppm) | |
| Protein | 6-8 g / dl (60,000-80,000 ppm) | 0.72-2.45 mg / ml (720-2,450 ppm) |
| Lipids | 35-135 mg / dl (350-1350 ppm) | 0.9-1.3 mg / dl (9-13 ppm) |
| DNA | 14-17 mg / l (14-17 ppm) | 1 – 100 ng / μ l (1-100 ppm) |
| RNA | 144-166 mg / l (144-166 ppm) | 4,912 – 15473 ng / μ l (4,912-15,473 ppm) |
| Saccharides | 80-120 mg / dl (800-1,200 ppm) | 0.005-0.01 mg / ml (5-10 ppm) |

4.5 Raman spectral data acquisition

Raman measurements for simulate samples, whole blood, saliva, pure biochemical components, and prostatic cell lines (PC3, PNT1a) were done with similar configuration. Briefly, the spectrograph was tuned at 600g / mm with 750 nm blazing wavelength. Actual spectral measurements were performed in the range of 393-2063 cm^{-1} , spectral resolution $\approx 1.35 \text{ cm}^{-1}$, exposure time = 120 seconds, using an 80x objective of the microscope. The measured laser power at sample surface and the spot size of excitation beam were $\approx 10.38 \pm 0.09 \text{ mW}$, and $\approx 51.85 \mu\text{m}$, respectively. The instrument was calibrated before fresh measurements using the reference band of silicon at 520.5 cm^{-1} , in 2 hours intervals to ensure wavelength stability (Desroches *et al.*, 2015). Automatic cosmic rays removal was done by the window based STR Raman software (version 1.9.3). To obtain a mean spectrum, 15-20 spectra were measured from five random points for each sample. Raman measurements were done in darkness to minimize possible spectral artifacts from fluorescent and microscope light sources (Desroches *et al.*, 2015).

Table 4.4. Calibration set design for biochemical components formulation in whole blood and saliva

| Sample | Concentration (mg / ml) | | | | | | |
|--------|-------------------------|-----------|-----------|-----------|-----------|-----------|--------------------|
| | Albumen | Glycogen | Glutamate | Glycerol | Glycine | RNA | Triolein |
| 1 | 0.5 | 0.1 | 0.0000001 | 0.01 | 0.2 | 0.0000001 | 0.001 |
| 2 | 0.4 | 0.3 | 0.000001 | 0.3 | 0.2 | 0.01 | 0.001 |
| 3 | 0.01 | 0.001 | 0.01 | 0.0001 | 0.3 | 0.4 | 0.00001 |
| 4 | 0.00001 | 0.1 | 0.2 | 0.3 | 0.2 | 0.4 | 0.01 |
| 5 | 0.000001 | 0.0000001 | 0.2 | 0.4 | 0.0000001 | 0.3 | 0.2 |
| 6 | 0.01 | 0.5 | 0.01 | 0.00001 | 0.0000001 | 0.01 | 0.00001 |
| 7 | 0.000001 | 0.0001 | 0.01 | 0.001 | 0.01 | 0.0001 | 0.1 |
| 8 | 0.1 | 0.0000001 | 0.2 | 0.01 | 0.00001 | 0.3 | 0.4 |
| 9 | 0.4 | 0.001 | 0.5 | 0.3 | 0.00001 | 0.001 | 0.01 |
| 10 | 0.5 | 0.4 | 0.2 | 0.3 | 0.4 | 0.01 | 1*10 ⁻⁷ |
| 11 | 0.0001 | 0.2 | 0.01 | 0.3 | 0.01 | 0.0000001 | 0.1 |
| 12 | 0.2 | 0.4 | 0.01 | 0.000001 | 0.5 | 0.5 | 0.01 |
| 13 | 0.001 | 0.001 | 0.3 | 0.3 | 0.2 | 0.5 | 1*10 ⁻⁶ |
| 14 | 0.5 | 0.01 | 0.5 | 0.3 | 0.2 | 0.001 | 0.1 |
| 15 | 0.001 | 0.2 | 0.001 | 0.5 | 0.000001 | 0.3 | 0.01 |
| 16 | 0.0000001 | 0.01 | 0.00001 | 0.0000001 | 0.4 | 0.000001 | 0.5 |
| 17 | 0.5 | 0.4 | 0.2 | 0.001 | 0.5 | 0.001 | 0.5 |
| 18 | 0.1 | 0.5 | 0.5 | 0.5 | 0.001 | 0.2 | 0.3 |
| 19 | 0.00001 | 0.01 | 0.0000001 | 0.00001 | 0.01 | 0.5 | 0.01 |
| 20 | 0.000001 | 0.01 | 0.0001 | 0.1 | 0.001 | 0.01 | 0.0001 |
| 21 | 0.01 | 0.00001 | 0.4 | 0.3 | 0.5 | 0.2 | 1*10 ⁻⁷ |
| 22 | 0.01 | 0.5 | 0.1 | 0.001 | 0.4 | 0.3 | 0.3 |
| 23 | 0.1 | 0.5 | 0.3 | 0.01 | 0.1 | 0.01 | 0.5 |
| 24 | 0.5 | 0.2 | 0.3 | 0.01 | 0.1 | 0.01 | 0.5 |
| 25 | 0.0000001 | 0.000001 | 0.00001 | 0.0001 | 0.001 | 0.5 | 0.01 |

4.6 Raman spectral analysis

In this study, prominent and subtle biochemical alterations in the studied samples were determined, followed by various machine learning techniques analysis procedures, as summarized in the conceptual framework in Figure 4.7. The spectral data analysis was restricted to the 500-1800 cm^{-1} range. The spectral data analysis was restricted to the 500–1800 cm^{-1} range. The spectral region 500-1800 cm^{-1} is the most frequently observed spectral region for Raman based biochemical investigations and is therefore considered the fingerprint region for biological specimens (Byrne *et al.*, 2015). In general, depending on the goal of the study, spectral data pre-processing involved co-adding, averaging, linear baseline correction, and spectral normalization (Bryne *et al.*, 2015; Ryabchykov *et al.*, 2019). Spectral analysis was done in two steps: first, by examining the prominent bands, followed by analysis of subtle spectral markers (weak variance signals).

4.6.1. Determination of prominent Raman bands for cancer diagnostics

The prominent Raman band alterations were investigated by comparing the band intensity profiles of control and diseased samples. To further investigate the band intensity differences in these spectral profiles, the difference spectra were computed by subtracting the normalized mean spectral intensities of normal samples from the normalized mean spectral intensities of diseased samples. The two sample *t*-test ($p < 0.05$) was used to determine statistical significance of difference bands.

4.6.2. Determination of subtle Raman bands for cancer diagnostics

The subtle Raman bands in prostate, breast cancer and leukemia progression were determined using a technique that has not been previously employed in other related leukemia and breast cancer studies (MartinEspinoza *et al.*, 2008; Babrah *et al.*, 2007; Feng *et al.*, 2015; Nargis *et al.*, 2019; Bilal *et al.*, 2017; Vargas-Obieta *et al.*, 2016; Cervo *et al.*, 2015; Moisoiu *et al.*, 2019; Yu *et al.*, 2017) i.e., the use of intermediate and high-order principal components. In this study, applicability of intermediate and high-order principal components in revealing subtle biochemical alterations during cancer progression was tested with prostatic cells spectral datasets.

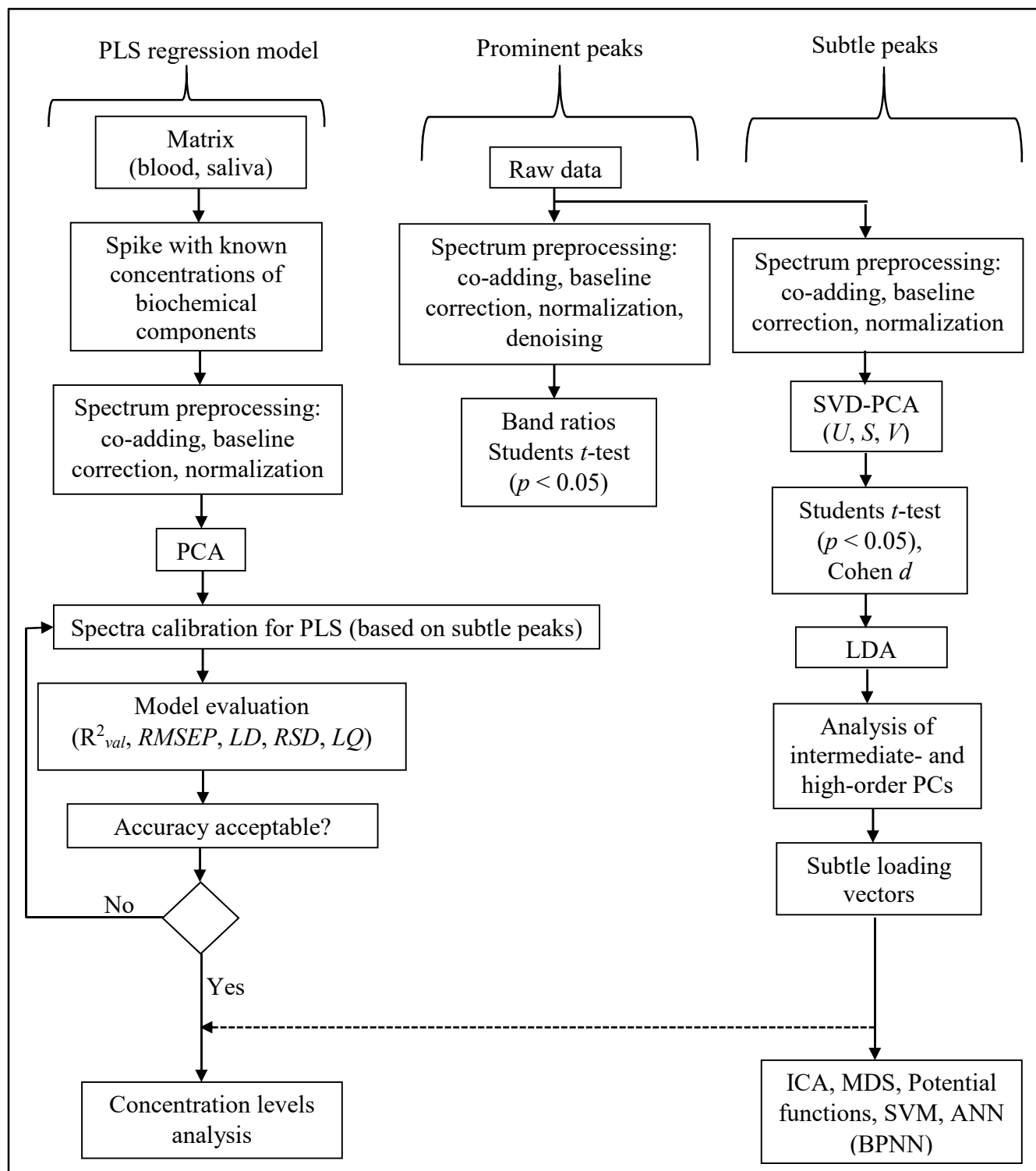


Figure 4.7 Schematic overview of the data processing and machine learning steps explored in this study. PLS, partial least squares; SVD, singular value decomposition; PCA, principal component analysis; R^2_{val} , root-mean-square error of cross-validation; $RMSEP$, root mean squared error of prediction; RSD , relative standard deviation; LD , limit of detection; LQ , limit of quantification; U , scores; S , size of scores; V , loadings; ICA, independent component analysis; MDS, multidimensional scaling; SVM, support vector machine; ANN, artificial neural networks; BPNN, backpropagation neural networks.

First, all collected raw spectra were baseline corrected followed by normalization. The prostatic cells data were normalized to their maximum intensity whereas the leukemia and breast cancer datasets were normalized at the CH₂ deformation band near 1445 cm⁻¹. The 1445 cm⁻¹ band was selected because it reveals biochemical information regarding protein to lipid ratios in biological samples (Movasaghi *et al.*, 2007). It should be noted, however, that spectral datasets were not denoised in order to retain all relevant features. Then, the Raman spectra of healthy (control) samples were combined with spectra of diseased samples to form single matrices i.e., $X_{w \times n}$ where w represented wavenumbers (781) and n represented total number of spectra (i.e., sum of controls and diseased samples' spectra). For prostatic cells, three matrices labelled stage₁ ($X_{781 \times 154}$), stage₂ ($X_{781 \times 160}$), and stage₃ ($X_{781 \times 198}$) were obtained. For breast cancer, the matrices for blood samples were grade₁ ($X_{781 \times 518}$), grade₂ ($X_{781 \times 573}$), and grade₃ ($X_{781 \times 671}$) whereas matrices for saliva samples were grade₁ ($X_{781 \times 559}$), grade₂ ($X_{781 \times 611}$), and grade₃ ($X_{781 \times 775}$). For leukemia, samples were successfully obtained from consenting patients whose malignancy was at advanced stages (stage 3). No consenting patient with stage 1 or stage 2 leukemia malignancy was available. Therefore, the matrices for leukemic blood and saliva samples were grade₃ ($X_{781 \times 408}$) and grade₃ ($X_{781 \times 421}$), respectively.

In order to understand the samples (scores) discrimination and their related biomarkers, all of these matrices were analyzed using SVD-PCA (Martinez *et al.*, 2005). With this procedure, three matrices; U representing scores, S representing scores' size, and V representing loadings were obtained, according to the following equation (Cordella, 2012):

$$M_{mn} = U_{mm} S_{mn} V_{nn}^T \quad (4.2)$$

where $U^T U = 1$; $V^T V = 1$. In this study, U , S , and V yield information regarding scores (samples) / spectra discrimination, number of principal components, and the correlation loadings (wavenumbers), respectively. This was coded in MATLAB 2018a scripting environment, as detailed in Appendix III. The number of PCs were determined by examining the scree plots. The statistical significance of PCs was calculated using the Students t -test and effect effect size (Pearson correlation coefficients (r), Cohen's d values) criteria (Sullivan *et al.*, 2012). The Cohen's d values for each PC scores were determined using the following equations (Thalheimer *et al.*, 2020):

$$d = \frac{(\bar{x})_t - (\bar{x})_c}{s_p} \quad (4.3)$$

$$s_p = \sqrt{\frac{(n_t - 1)s_t^2 + (n_c - 1)s_c^2}{n_t + n_c}} \quad (4.4)$$

where d = Cohen's d effect size, (\bar{x}) = mean of diseased or normal conditions, s = standard deviation, n = number of samples, t = diseased condition, c = control condition. The Cohen d effect sizes were classified as small ($d = 0.2$), medium ($d = 0.5$), and large ($d \geq 0.8$) (Sullivan *et al.*, 2012). This was done in Microsoft Excel (2010) platform. The resulting p -values were adjusted using the Holm–Bonferroni method to ensure that all experiments had a family-wise alpha of 0.05 (Jafari & Ansari-pour, 2019). This was done using $p.adjust$ (Bonferoni) function in R scripting environment. Using a variety of adjustment methods, the $p.adjust$ (Bonferoni) function calculates modified p -values from a collection of unadjusted p -values.

Next, LDA classification models were created using the fitting coefficients of principal components from PCA, and the classifier was trained using a k -fold cross validation process. The correlation between the original discrete variables and PC scores was evaluated by examining the computed loadings. The loadings were used to approximate the biochemical details that the PC scores and the original discrete variables shared in this analysis. The scatter plots of intermediate- and higher-order principal components were further evaluated and the subtle spectral markers (loading vectors) extracted by visual examination for further analysis.

4.6.3 Quantitative Raman spectral analysis using partial least-squares regression

The partial least squares (PLS) regression model was used to match the mean spectra of the calibration samples to the mean spectra of the different pathologies in whole blood and saliva, allowing for quantitative spectrochemical analysis of subtle band alterations (Høy *et al.*, 2012). Cross-validation based on the singular value decomposition PCA approach was used to perform partial squares regression. PCA served to reduce dimensionality of spectral datasets. About 70% ($\approx 2/3$) of spectra were used for training / calibrating the model (tuning the parameters of a model) and the remaining 30% ($\approx 1/3$) of spectra were used for testing the model (evaluating the model's performance). The data were mean-centered, and the internal validation was done using k ($=10$) fold cross-validation. The training set was randomly partitioned into k ($=10$) equal-sized

subsamples using this technique. A single subsample from the k (=10) subsamples was held as the validation set for testing the model, while the remaining $k-1$ (=9) subsamples served as the training set. This procedure was repeated k (=10) times, with each subset serving as the validating set once. Then, the performance of the model was determined by calculating the average performance across all k (=10) trials. The determined subtle peaks were used for subsequent spectral calibration and model evaluation. The accuracy was optimized by careful choice of optimal components during cross-validation process. The non-negative constrained fitting method was chosen in this study to eliminate the possibility of negative coefficients, which would result in distorted fitting.

At a significance level of 5%, outliers were identified by detecting i) residual levels in the analytical concentrations or, ii) samples with high leverage and residuals in the spectral data (Valderrama *et al.*, 2007). Correlating the real values with the expected values from the prediction set was used to assess the PLS model's fit. In addition, the linearity was used to estimate the model linearity qualitatively using the residuals distribution against parameters of reference (Frost, 2016). The coefficient of multiple determination during cross-validation (R^2_{val}) and the root mean squared error of prediction ($RMSEP$) were used to determine the model's accuracy (Frost, 2016; Sichangi *et al.*, 2018).

$$RMSEP = \sqrt{\frac{\sum_{i=1}^n (y_{pred,i} - y_i)^2}{n}} \quad (4.5)$$

where $y_{pred,i}$ refers to predicted concentration; y_i is the known concentration for sample i ; n represents samples count. The limits of detection (LD) and quantification (LQ) were determined according to the equation 4.6 and equation 4.7, respectively (Desimoni *et al.*, 2015):

$$LD = 3.3 * \left(\frac{\delta_Y}{b} \right) \quad (4.6)$$

$$LQ = 10 * \left(\frac{\delta_Y}{b} \right) \quad (4.7)$$

where $\delta_{Y/X}$ represents the standard deviation of the response (δ) and b represents the slope of the calibration curve. The values 3.3 and 10 are expansion factors obtained assuming a 95% confidence level (Desimoni *et al.*, 2015).

The accuracy and reliability of PLS regression model was evaluated by analysis of separately prepared standard whole blood and saliva simulates spiked with prepared concentrations

of various biochemical components: albumen-0.4 mg/ml; glycogen-0.1 mg/ml; glutamate-0.001 mg/ml; glycerol-0.01 mg/ml; RNA-0.002 mg/ml and triolein-0.3. At each step, four measurements were taken, and the standard deviation of the measurements was calculated using the equation below (Gontijo *et al.*, 2014),

$$\sigma_i = \sqrt{\frac{\sum_{j=1}^{r_i} (\hat{y}_{ij} - \tilde{y}_i)^2}{r_i - 1}} \quad (4.8)$$

where \tilde{y}_i represents the mean value of the measurements per level and r_i represented the total number of measurements made on each level. The standard deviation relative to the mean was calculated as follows (Gontijo *et al.*, 2014),

$$RSD = \frac{\sum_{i=1}^n \frac{\sigma_i}{\tilde{y}_i} \cdot 100}{p} \quad (4.9)$$

where σ_i and p are standard deviation and the number of concentration levels evaluated, respectively.

4.6.4 Multivariate statistical analysis of trace biomarkers alterations

In a Raman spectrum, the different spectral markers can be regarded as independent variables, and the intensities correspond to the magnitudes of the variables. Therefore, the matrices of the determined subtle spectral markers were subjected to the independent component analysis (ICA). The ICA was adopted on the fast fixed-point estimation algorithm using Maximum Likelihood (ML) criterion in MATLAB 2018a scripting environment, as detailed in Appendix D. The use of ICA analysis was motivated by the fact that Raman data consists of a set of independent signals (e.g. various forms of proteins) additively combined to form a single protein band, which necessitated need for mutual independence during quantification. In ICA, spectral data were mean centered, whitened and followed by several iterations until convergence leading to determination of independent components. The algorithm was based on the following expression (Hyvärinen *et al.*, 2000).

$$W^+ = W + \text{diag}(\alpha_i)[\text{diag}(\beta_i) + E\{g(y)y^T\}]W \quad (4.10)$$

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 $\text{diag}(\alpha_i)$ and $\text{diag}(\beta_i)$.

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 version of the fixed-point algorithm was used to ensure algorithm convergence. The maximum number of iterations were set at 1000.

A challenge with ICA is the assumption that independent signals $S = [S_1(t), S_2(t), \dots, S_N(t)]$ combine linearly to form signals $Y(t) = AS(t)$, where both A (mixing matrix) and S are unknown. It's worth noting that linear transform methods enforce a linear structure on the analyzed data and can't model nonlinearity well (Wang *et al.*, 2014). For modeling high nonlinear dimensional data sets with a limited number of samples, a nonlinear feature extraction approach such as multidimensional scaling (MDS) would be most suitable (Wang *et al.*, 2014). The aim of MDS is to find a configuration of data points in a low-dimensional space such that the distances between points in the low-dimensional space accurately reflect the proximity between objects in the full-dimensional space (Martinez *et al.*, 2005). In the present study, the ICA analysis was extended to include MDS as a potential non-linear dimensional reduction algorithm in blood Raman datasets. To achieve excellent diagnostic sensitivity, independent components were further analyzed using Minkowski MDS metrics (Weinberg, 1991), PLS-DA, Mahalanobis multidimensional metrics (MDS) and the potential functions (kernel density estimators). This was done in MATLAB 2018a scripting environment, as detailed in Appendices V to VII.

The Mahalanobis distance calculation has a well-known mathematical foundation (McLachlan, 1999), and it is commonly used for spectral discrimination. It provides a statistical measure of how well the unknown sample spectrum fits or does not match in addition to spectral discrimination (Chowdary *et al.*, 2006). As a result, it is a statistical measure of the distance between two spectra. Previous works have demonstrated that kernel density estimators (potential functions) are always as good as discrete functions and often better (Coomans *et al.*, 1981), and their utility can be considered as one method of optimizing the ICA-MDS model to achieve higher diagnostic accuracy. Briefly, each sample of the training set is treated as a point in the pattern space by the potential functions. There is a potential field around this point that decreases with distance from the sample (Coomans *et al.*, 1981). As a result, samples in close proximity to a strong potential field appear to cluster in space. In the present study, the total potential of the class in the location of the test sample was used to evaluate the classification of a new sample into one of the classes (controls versus diseased). The cumulative potential was calculated by adding the

individual potentials of the class samples and then classifying the test item into the class with the highest cumulative potential (Forina *et al.*, 1991).

Separately, diagnostic models based on SVM and ANN were used to detect and characterize breast and leukemia cancers spectra during classification. For ANN, we used the backpropagation neural networks (BPNN) algorithm. Detailed implementations of BPNN and SVM models in MATLAB 2018a scripting environment are provided in Appendices H and I, respectively. For SVM, the fitting function defined in equation (3.25) was used. The parameter α_i was restricted to $0 \leq \alpha_i \leq C$ and estimated by maximizing a Lagrangian. The linear and radial basis functions (RBF) SVM kernel functions defined in equations (3.26) and (3.27) were adopted. The data analysis looked for the best RBF 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, k (=10) folds cross-validation was used during data analysis.

The backpropagation nets were implemented using the fitting function defined in equation (3.28), and performance estimated through the mean squared error (MSE) defined by equation (3.29). The adjustment of weights during training provides the adaptive fitting capabilities of backdrop nets (Marini *et al.*, 2008). To improve the BPNN model accuracy, learning was accomplished by adjusting the weights shown in equation (4.14), using error feedback from the training examples, so as to bring the network estimates of the correct outputs for the training patterns closer to the true values. The learning law was regulated by the learning rate (θ) (in the range of 0.0001 to 0.1) and the momentum term (α) (in steps of 0.1, range: 0.1-0.9). The learning rate (θ) was a user defined criterion (value) that enabled some control over the scale of the weight changes during training of the model, while the momentum term (α) governed the canceling of opposing components of the phase at successive positions and the enhancement of reinforcing components (Marini *et al.*, 2008). This allowed acceleration over long stretches of shallow but relatively steady gradient, as well as exit from local minima (Wythoff, 1993). The 10 folds cross-validation at 1000 number of iterations was applied in order to improve the model accuracy. About 15 to 30 neurons per layer were chosen, since they were observed to deliver the best accuracy

without over-prediction. Detailed implementations of BPNN and SVM models in MATLAB 2018a scripting environment are provided in Appendices VIII and IX, respectively.

The average performance of the PCA-LDA and PLS-DA models was assessed by calculating the sensitivity, specificity, and accuracy values. In this context, sensitivity characterized the test method's ability to detect the disease in diseased subjects, whereas specificity characterized the test method's ability to detect the absence of disease in healthy subjects (Vargas-Obieta *et al.*, 2016). The accuracy value represented the proportion of true positive results (both true positive and true negative) in the selected population. The sensitivity, specificity, and accuracy values were calculated as follows (Wang *et al.*, 2008):

$$\text{Sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \quad (4.11)$$

$$\text{Specificity} = \frac{\text{TN}}{(\text{TN} + \text{FP})} \quad (4.12)$$

$$\text{Accuracy} = \frac{\text{TN} + \text{TP}}{(\text{TN} + \text{FP} + \text{FN} + \text{FP})} \quad (4.13)$$

where TP, TN, FN, and FP represented the true positives, true negatives, false negatives, and false positives, respectively.

Chapter 5 Results and Discussions

As clearly described in section 4.2.1, a spectrometric analysis study was undertaken with the aim of evaluating the potential of intermediate- and high-order principal components in revealing subtle biochemical changes in spectral profiles of biological samples. Therefore, an *in vitro* tissue model based on metastatic (PC3) and normal (PNT1a) human prostate cell lines was chosen for this study. Based on the promising results obtained with prostatic cells datasets, we employed similar technique i.e., use of intermediate- and high-order principal components, on blood and saliva Raman spectra taken from healthy volunteers, breast cancer and leukemia patients.

5.1 Raman spectroscopic characterization of PC3 and PNT1a cells

5.1.1 Analysis of prominent biochemical alterations

Figure 5.1 (a) and (b) shows optical photomicrographs of unstained PC3 and PNT1a monolayer cells grown on calcium fluoride (CaF_2) substrates, respectively. As expected, the PC3 and PNT1a are typically adherent cells that grow attached to a substrate in discrete patches, usually with regular dimensions (ideally polygonal shapes). Here, it is noted the attached monolayer cells have epithelial morphological features with the nucleus (i), cytoplasm (ii), and cell wall (iii) locations discernible in both images.

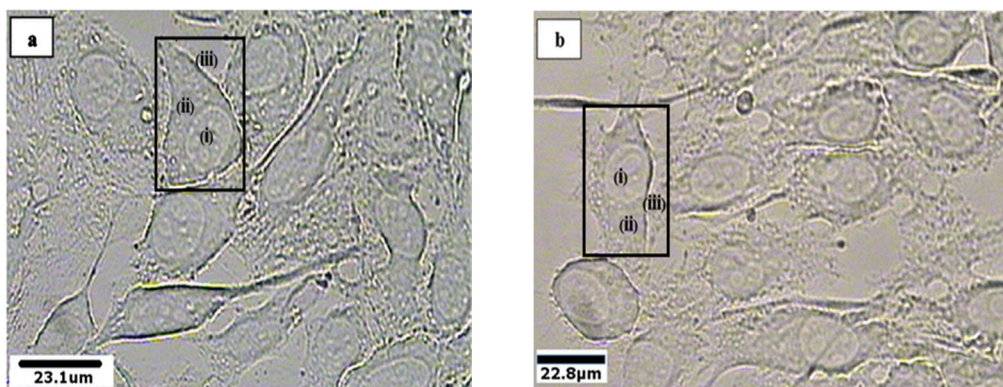
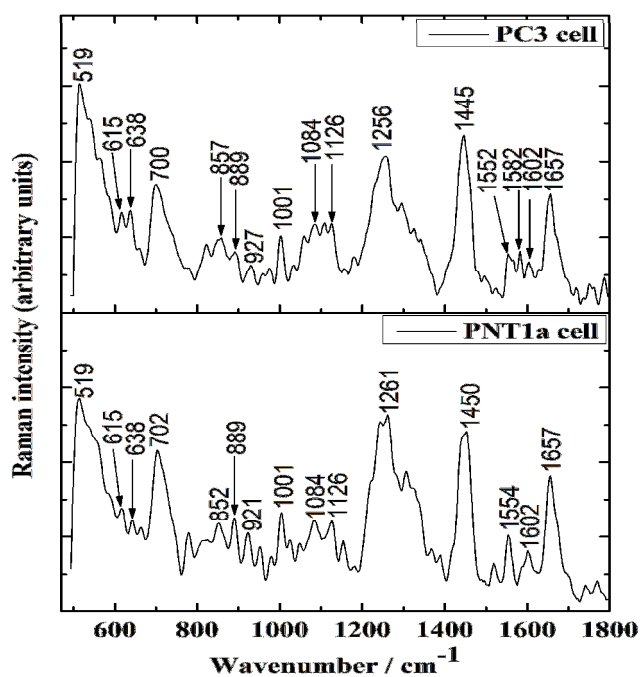


Figure 5.1 Photomicrographs of (a) PC3 and (b) PNT1a monolayer cells grown on calcium fluoride (CaF_2) substrates at 50x magnification, with (i) nucleus, (ii) cytoplasm and (iii) cell wall constituents clearly visible in both cells.

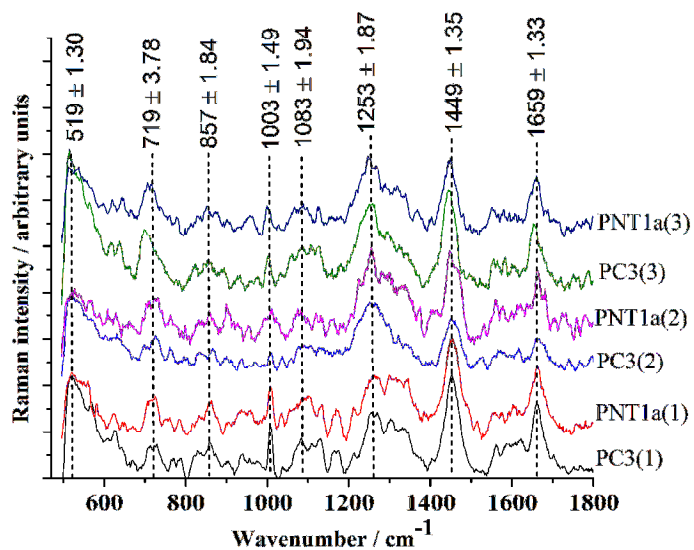
The biochemical assignments of peaks were done in accordance with the Raman spectroscopy of tissues, body fluids, or bio-molecules as reported in the literature (Chandra *et al.*, 2015; Movasaghi *et al.*, 2007; Gelder *et al.*, 2007). This was done in consideration of position and possible wavelength differences of each particular Raman band. Theoretically, spectral resolution in a dispersive Raman spectrometer is determined by many factors which include, spectrometer focal length, diffraction grating, laser wavelength and the detector (McCreely, 2001). Therefore, the Raman bands within $\pm 10 \text{ cm}^{-1}$ were considered to represent the same Raman peak due to potential of varying detection conditions and experimental errors. This choice of wavelength difference was further motivated based on the work of McCreely (L, 2001) who suggested that most analytical Raman applications involve liquids and solids in which Raman bandwidths are significantly greater than those in the gas phase, hence the narrowest linewidths encountered in most liquid and solid samples range from 3 cm^{-1} to 10 cm^{-1} (McCreely, 2001).

Figure 5.2 (a) shows an example of as-collected averaged, minimally denoised, and baseline corrected raw spectra in PC3 and PNT1a samples. Examination of both spectra show the primary Raman bands featuring at 519 cm^{-1} , 615 cm^{-1} , 638 cm^{-1} , $(700, 702 \text{ cm}^{-1})$, $(852, 857 \text{ cm}^{-1})$, 889 cm^{-1} , $(921, 927 \text{ cm}^{-1})$, 1001 cm^{-1} , 1084 cm^{-1} , 1126 cm^{-1} , $(1256, 1261 \text{ cm}^{-1})$, $(1445, 1450 \text{ cm}^{-1})$, $(1552, 1554 \text{ cm}^{-1})$, 1602 cm^{-1} and 1657 cm^{-1} . The respective Raman band assignments are provided in Table 5.1. Figure 5.2 (b) shows the stacked mean Raman spectrums that explain biochemical alterations occurring during all stages of cell proliferation, in both cell lines. It should be noted the spectra have been linearly offset for comparison purposes.

The spectra demonstrate a similar spectral pattern though there are minor Raman shifts. Figure 5.2(b) reveals the mean positions (and the standard errors) of notable major Raman band contributions in both cells, whose biochemical assignments are well known (Movasaghi *et al.*, 2007). They include $519 \pm 1.30 \text{ cm}^{-1}$ (phosphatidylinositol), $719 \pm 3.78 \text{ cm}^{-1}$ (phospholipids, nucleic acids), $857 \pm 1.84 \text{ cm}^{-1}$ (proline, tyrosine proteins), $1003 \pm 1.49 \text{ cm}^{-1}$ (phenylalanine), $1083 \pm 1.94 \text{ cm}^{-1}$ (C-N stretch of proteins and lipids), $1253 \pm 1.87 \text{ cm}^{-1}$ (nucleic acids, lipids), $1449 \pm 1.35 \text{ cm}^{-1}$ (C-H vibration of proteins and lipids) and $1659 \pm 1.33 \text{ cm}^{-1}$ (Amide I), with the strongest Raman bands occurring around $519 \pm 1.30 \text{ cm}^{-1}$, $1253 \pm 1.87 \text{ cm}^{-1}$, $1449 \pm 1.35 \text{ cm}^{-1}$ and $1659 \pm 1.33 \text{ cm}^{-1}$.



(a)



(b)

Figure 5.2 (a) Examples of as-collected averaged, minimally denoised, and baseline corrected raw spectra in PC3 and PNT1a cells, and (b) the mean spectra of cell samples. The spectra shown in (b) have been linearly offset for comparison. The numbers (1, 2, 3) enclosed in brackets identify stage 1 (48 hours), stage 2 (72 hours), and stage 3 (96 hours) spectral measurements.

Table 5.1 Raman band assignments of PC3 and PNT1a prostatic cells

| Raman shift (cm ⁻¹) | Functional groups and molecular vibration assignments | References |
|------------------------------------|---|--|
| 519 | $\delta(\text{CH}_2)$, $\delta(\text{CH}_3)$ deformations (phosphatidylinositol) | (Movasaghi <i>et al.</i> , 2007) |
| 615 | Cholesterol esters | (Chandra <i>et al.</i> , 2015) |
| 638 | C-C stretch, C-C twisting of tyrosine proteins | (Gelder <i>et al.</i> , 2007) |
| 700, 702 | $\nu(\text{C-S})$ stretch of cholesterol esters | (Chandra <i>et al.</i> , 2015) |
| 852, 857 | Ring breathing modes (tyrosine proteins) | (Jr <i>et al.</i> , 2014) |
| 889 | Structural protein modes of tumors | (Movasaghi <i>et al.</i> , 2007) |
| 921 | C-C stretch (proline ring, glucose, lactic acid) | (Chandra <i>et al.</i> , 2015) |
| 927 | $\nu(\text{C-C})$ stretch of proline and valine proteins | (Gelder <i>et al.</i> , 2007) |
| 1001 | Symmetric ring breathing mode of phenylalanine | (Magalhães <i>et al.</i> , 2018) |
| 1084 | Phosphodiester groups in nucleic acids, C-N stretching of proteins | (Chandra <i>et al.</i> , 2015) |
| 1126 | C-N stretch (proteins), $\nu(\text{C-C})$ stretch (phospholipids) | (Magalhães <i>et al.</i> , 2018) |
| 1256, 1261 | C-H bend (phospholipids), C-N stretch (proteins), CH ₂ twisting and wagging (lipids, proteins) CH ₂ bending (proteins) | (Jr <i>et al.</i> , 2014) |
| 1445 | CH ₂ CH ₃ bending modes of collagen and phospholipids, $\delta(\text{CH}_2)$, $\delta(\text{CH}_3)$ of collagen protein assignments, $\delta(\text{CH}_2)$, $\delta(\text{CH}_3)$ scissoring of phospholipids | (Corsetti <i>et al.</i> , 2018) |
| 1450 | CH ₂ bending (proteins), C-H vibration (lipids), Ring breathing modes of DNA / RNA bases | (Corsetti <i>et al.</i> , 2018) (Jr <i>et al.</i> , 2014) |
| 1552, 1554 | $\nu(\text{C=C})$ stretching modes (tryptophan), Amide II (| (Huang <i>et al.</i> , 2003) |
| 1582 | C = C bending mode of phenylalanine (proteins) | (Corsetti <i>et al.</i> , 2018) |
| 1602 | C=C bending modes (phenylalanine, tyrosine) | (Corsetti <i>et al.</i> , 2018) |
| 1657 | C=O stretch (proteins, lipids), C = C stretch (lipids), nucleic acids | (Chandra <i>et al.</i> , 2015) (Jr <i>et al.</i> , 2014) |

The difference spectra between the normalized mean baseline corrected spectrum of PC3 cells and that of PNT1a cells are shown in Figure 5.3(a)–(c). Positive bands explain alterations that were more prevalent in PC3 cells, while negative bands explain alterations that were more prevalent in PNT1a cells.

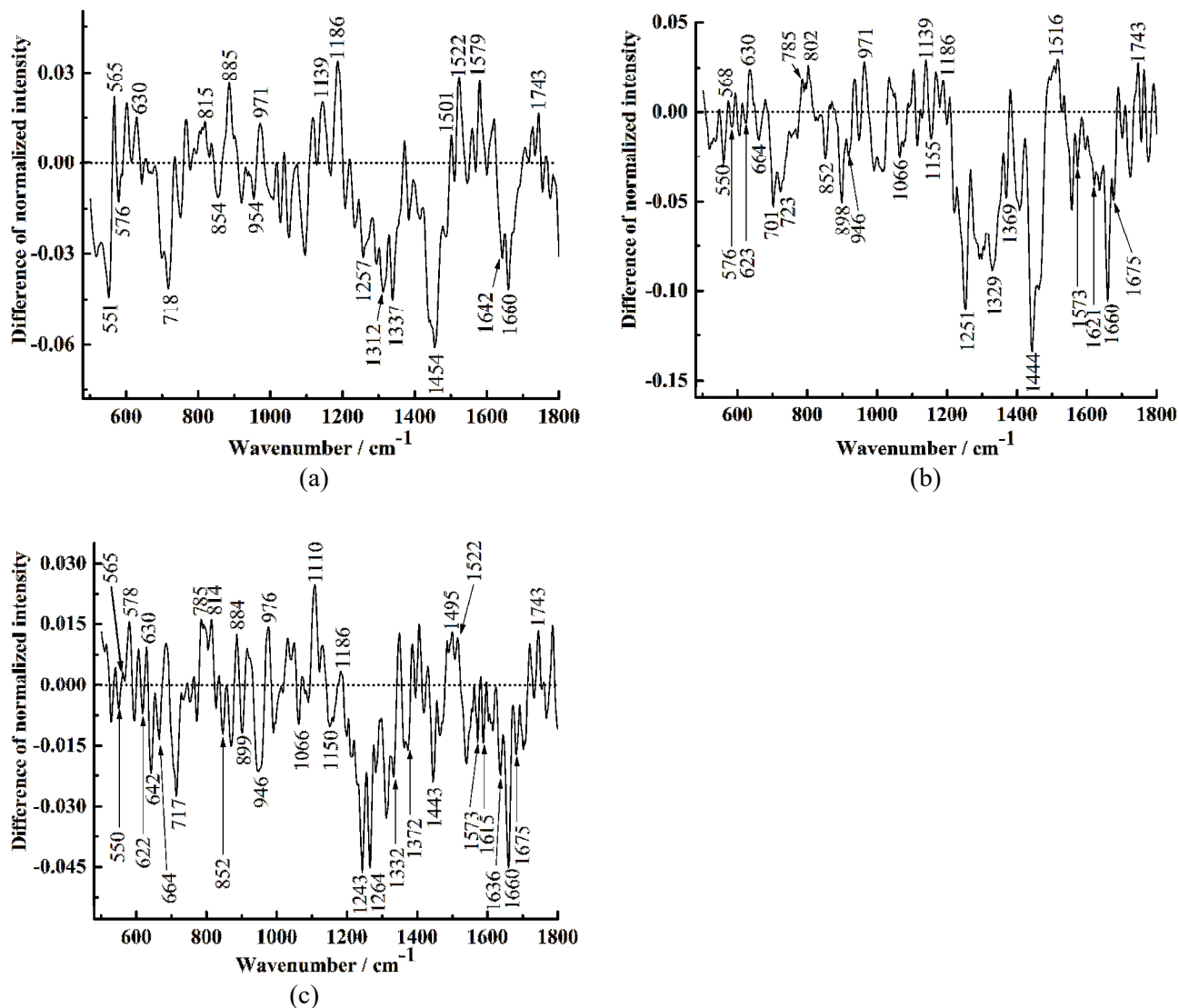


Figure 5.3 The difference spectrum between the normalized (a) stage 1 (48 hours), (b) stage 2 (72 hours), and (c) stage 3 (96 hours) PC3 and PNT1a spectral datasets. Examination of difference spectrum according to the cell proliferation cycles shows PC3 samples have heightened Raman peaks at $566 \pm 0.70 \text{ cm}^{-1}$, 630 cm^{-1} , $972 \pm 1.17 \text{ cm}^{-1}$, 1186 cm^{-1} , $1520 \pm 1.41 \text{ cm}^{-1}$, and 1743 cm^{-1} . The PNT1a samples have heightened Raman peaks at $550 \pm 0.23 \text{ cm}^{-1}$, $719 \pm 1.31 \text{ cm}^{-1}$, $852 \pm 0.47 \text{ cm}^{-1}$, $948 \pm 1.88 \text{ cm}^{-1}$, $1250 \pm 2.86 \text{ cm}^{-1}$, $1332 \pm 1.64 \text{ cm}^{-1}$, $1450 \pm 2.20 \text{ cm}^{-1}$, and 1660 cm^{-1} .

Generally, the difference spectra shows PC3 samples exhibiting heightened alterations around $566 \pm 0.70 \text{ cm}^{-1}$ (cytosine, guanine), 630 cm^{-1} (glycerol), $972 \pm 1.17 \text{ cm}^{-1}$ (nucleic acids), 1186 cm^{-1} (guanine, anti-symmetric phosphate vibrations), $1520 \pm 1.41 \text{ cm}^{-1}$ (cytosine, C-C stretch mode (β – carotene accumulation)) and 1743 cm^{-1} (C = O stretch mode, lipids) during all stages of proliferation cycle (Chandra *et al.*, 2015; Jr *et al.*, 2014). Similarly, PNT1a samples have same heightened biochemical alterations around $550 \pm 0.23 \text{ cm}^{-1}$ (cytosine, guanine, tryptophan, glycerol), $719 \pm 1.31 \text{ cm}^{-1}$ (nucleic acids), $852 \pm 0.47 \text{ cm}^{-1}$ (proline, tyrosine, polysaccharides), $948 \pm 1.88 \text{ cm}^{-1}$ (proline), $1250 \pm 2.86 \text{ cm}^{-1}$ (Amide III), $1332 \pm 1.64 \text{ cm}^{-1}$ (nucleic acids, CH_3CH_2 wagging (of collagen)) $1450 \pm 2.20 \text{ cm}^{-1}$ (lipids, proteins) and 1660 cm^{-1} (Amide I) during all stages of proliferation cycle (Movasaghi *et al.*, 2007; Gelder *et al.*, 2007). It is also noted stage 2 and stage 3 PNT1a spectra have common band alterations around 623 cm^{-1} (phenylalanine, adenine), 664 cm^{-1} (nucleic acids), $898 \pm 0.20 \text{ cm}^{-1}$ (saccharides), 1066 cm^{-1} (proline collagen), $1152 \pm 1.44 \text{ cm}^{-1}$ (proteins, carotenoids), $1370 \pm 0.86 \text{ cm}^{-1}$ (saccharides), 1573 cm^{-1} (guanine, adenine, proteins), $1618 \pm 1.73 \text{ cm}^{-1}$ (tryptophan) and 1675 cm^{-1} (Amide I) (Movasaghi *et al.*, 2007; Gelder *et al.*, 2007). Further, the spectral marker at 576 cm^{-1} (phosphatidylinositol) is present during the first two stages of PNT1a cell proliferation cycles, and later present during late proliferation cycle (stage 3) of malignant (PC3) cells. This suggests presence of enhanced lipid alterations during advanced malignancy.

To evaluate whether the observed bands (Figure 5.3) could be utilized in prostate cancer diagnosis, we examined their respective statistical significance (based on their actual preprocessed mean intensities) using a combination of the Student *t*-test and ANOVA test: two factor without replication tools in Microsoft[®] Excel. The Student *t*-test for each band were determined using array of matrices from intensities arising from the the mean \pm standard deviations of the bands of interest. The tests on averaged intensities around these bands showed they were statistically significant ($p < 0.05$), with exception of 576 cm^{-1} band, meaning the biochemical changes at the 576 cm^{-1} band could not be used to segregate the control (PNT1a) group from the diseased (PC3) cells group.

The peak intensity ratios of Raman spectra measurements have been earlier reportedly utilized in classifying diseased and healthy samples (Huang *et al.*, 2003). A similar analysis was performed on the observed prominent difference bands in Figure 5.3 (a) – (c), where ratio values (i.e. I_C / I_N) were calculated by dividing the normalized intensities of PC3 cells spectra (I_C) by normalized intensities of PNT1a cells spectra (I_N). Table 5.2 highlights the bands ($566 \pm 0.70 \text{ cm}^{-1}$, 630 cm^{-1} , $1370 \pm 0.86 \text{ cm}^{-1}$, $1618 \pm 1.73 \text{ cm}^{-1}$) whose band intensity ratios were found to increase

or decrease with stage of cell proliferation. The ratio values at $566 \pm 0.70 \text{ cm}^{-1}$ and 630 cm^{-1} bands increased with stage of cell proliferation, while ratio values at $1370 \pm 0.86 \text{ cm}^{-1}$ and $1618 \pm 1.73 \text{ cm}^{-1}$ were found to decrease with stage of cell development.

Table 5.2. Comparison of peak intensity ratios between malignant (PC3) and normal cells (PNT1a) at $566 \pm 0.70 \text{ cm}^{-1}$, 630 cm^{-1} , $1370 \pm 0.86 \text{ cm}^{-1}$, and $1618 \pm 1.73 \text{ cm}^{-1}$ spectral regions

| Raman shift (cm^{-1}) | Stage 1 | | Stage 2 | | Stage 3 | | Band ratios | | | |
|-------------------------------------|-------------------------------------|------|---------|------|-------------|------|--|--|--|--|
| | Raman intensities (arbitrary units) | | | | Band ratios | | | | | |
| | PNT1a | PC3 | PNT1a | PC3 | PNT1a | PC3 | PC3(1) $\overline{\text{PNT1a(1)}}$ | PC3(2) $\overline{\text{PNT1a(2)}}$ | PC3(3) $\overline{\text{PNT1a(3)}}$ | |
| 565 | 0.90 | 0.87 | 0.75 | 0.85 | 0.75 | 0.87 | 0.967 | 1.133 | 1.16 | |
| 568 | 0.81 | 0.87 | 0.77 | 0.86 | 0.77 | 0.89 | 1.074 | 1.117 | 1.156 | |
| 630 | 0.69 | 0.79 | 0.73 | 0.86 | 0.62 | 0.85 | 1.145 | 1.178 | 1.371 | |
| 1368 | 0.51 | 0.47 | 0.72 | 0.66 | 0.68 | 0.58 | 0.922 | 0.917 | 0.853 | |
| 1372 | 0.40 | 0.44 | 0.71 | 0.65 | 0.60 | 0.53 | 1.1 | 0.915 | 0.883 | |
| 1617 | 0.32 | 0.35 | 0.46 | 0.43 | 0.58 | 0.53 | 1.094 | 0.935 | 0.914 | |
| 1621 | 0.30 | 0.32 | 0.59 | 0.54 | 0.49 | 0.44 | 1.067 | 0.915 | 0.898 | |

The increasing peak ratios around $566 \pm 0.70 \text{ cm}^{-1}$ band suggests increase of nucleic acids bases (cytosine, guanines) with malignancy. This correlates with other closely related prostatic studies (Stone *et al.*, 2007; Crow *et al.*, 2003; Taleb *et al.*, 2006) which have suggested that spectral intensities of DNA related bands in prostate samples increase with malignancy, a factor attributed to enlarged nuclei in malignant cells than for normal cells and therefore greater abundance of DNA content in malignant samples (Taleb *et al.*, 2006). For instance, findings by Crow *et al.*, (2003) showed nucleic acid contents in malignant prostate biopsies were higher compared to benign prostatic hyperplasia (BPH) ones. Further, a study that investigated biochemical alterations for the different tissue pathologies within the bladder and prostate gland region noted the DNA content increased with malignancy (Stone *et al.*, 2007), and DNA contents were observed to be higher in malignant (LNCaP) cells when compared with normal (PNT1a) cells (Taleb *et al.*, 2006). The increasing peak ratios around 630 cm^{-1} band suggests increment of lipids with malignancy. A closely related study by Matias *et al.*, (2011), observed adipocyte levels increase with severity of malignancy, a factor attributed to the production of lipids via *de novo* lipogenesis (Long *et al.*,

2018). The decreasing peak ratio values observed around $1370 \pm 0.86 \text{ cm}^{-1}$ and $1618 \pm 1.73 \text{ cm}^{-1}$ bands suggests that saccharides and tryptophan levels decreased with malignancy. The decrease of saccharides levels fit in with the findings of a previous cell line based - study where glycogen content was found to be lower in prostatic adenocarcinoma pathologies, compared to benign pathologies (Crow *et al.*, 2003), a factor attributed to enhanced glucose uptake by cells during onset of tumor development for conversion to lactate molecules necessary for energy production during cell proliferation (Klement *et al.*, 2013). The systematic decrement of tryptophan levels with malignancy is an indication of prostate malignancy undergoing increased tryptophan degradation.

PCA is a suitable dimensional reduction multivariate technique that filter out variables according their levels of variance (Corsetti *et al.*, 2018). PCA was applied to the observed difference bands' spectra (Figure 5.3 a-c) for stage 1, stage 2 and stage 3 of proliferation cycle. The contribution rates to the total variation of spectra of the first two PCs (PC 1 and PC 2) were 99.13%, 99.19%, and 99.64%, respectively; hence, the first two principal components accounted for a greater percentage of the total variability, and were therefore selected for subsequent analysis. The score plot of the control and diseased groups was achieved using PC 1 and PC 2; which represented group discrimination. Further, LDA was used on PC 1 and PC 2 to evaluate their ability of spectra discrimination. In addition, LDA was included to ensure that the distribution of the data across the scatter plot was due to variations in spectral features correlated to a pathological state but not other less relevant biological parameters e.g. spectral noise (Nargis *et al.*, 2019). It was noted that majority the scores / spectra belonged to PNT1a or PC3 classes separated at the two sides of the discriminant line (Figure 5.4 a-c), demonstrating the samples could be discriminated well using PCA-LDA. Diagnostic performance of PCA-LDA was based on accuracy, sensitivity and specificity parameters, as explained in equations 4.11 to 4.13. The overall classification accuracies obtained were 97%, 98% and 98% for stages 1, 2 and 3 spectral datasets, respectively. The achieved sensitivity parameters were 100%, 100%, and 100%, respectively.

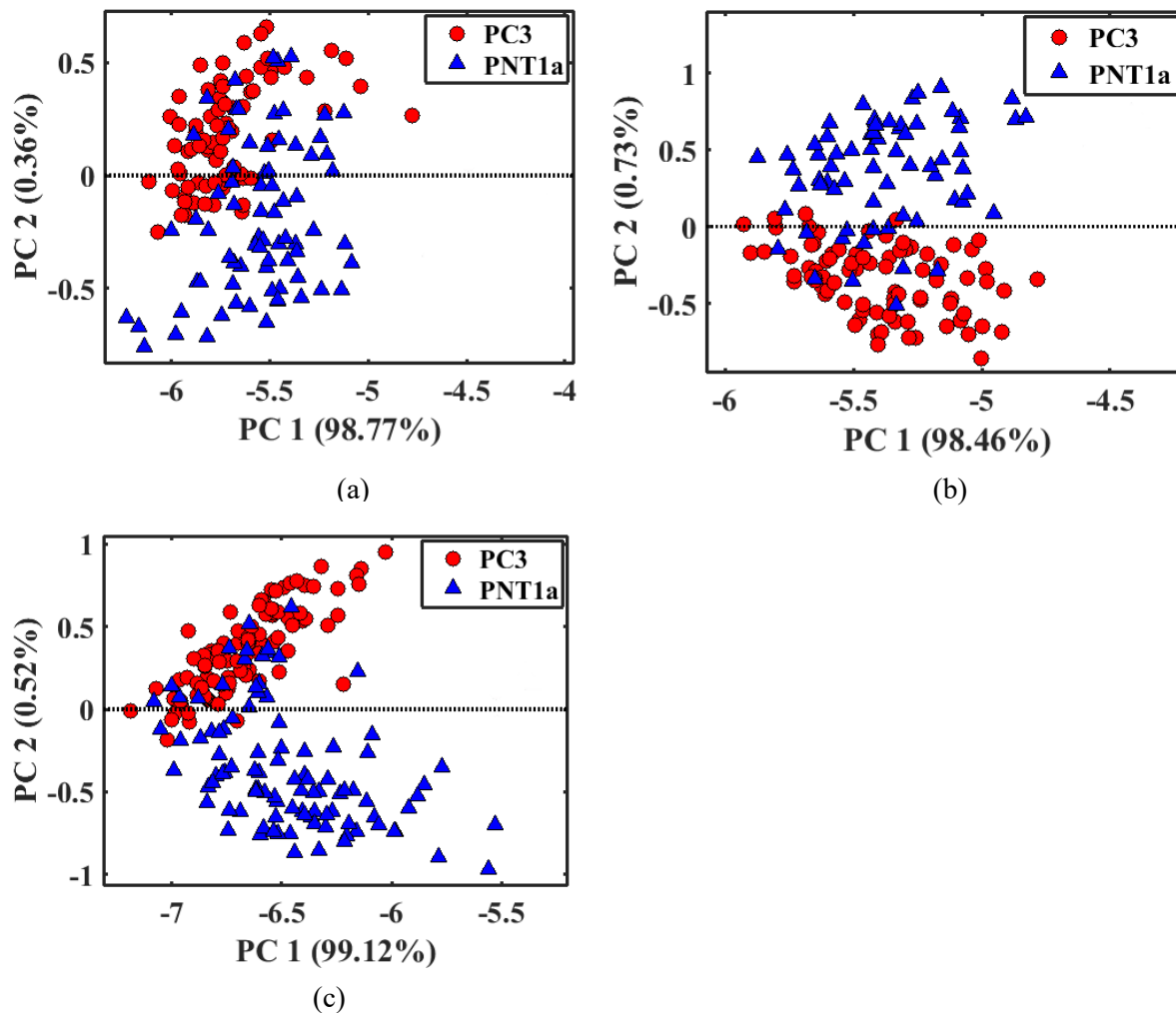
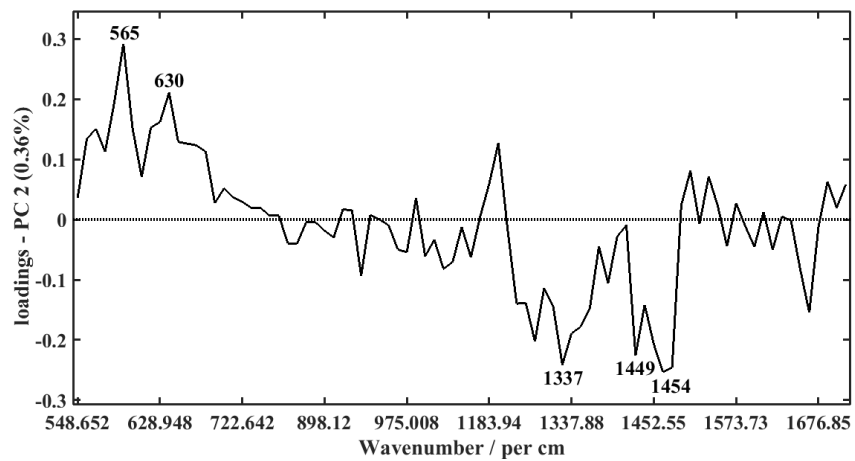
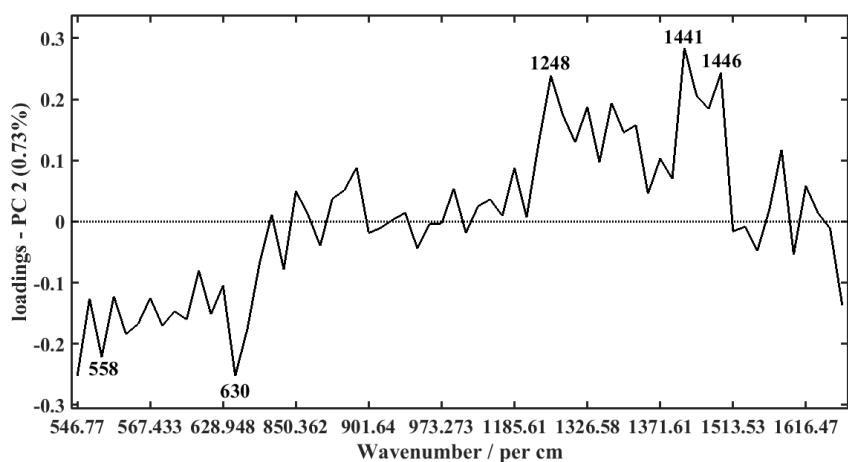


Figure 5.4 The score plots of PC 1 and PC 2 for (a) stage 1 (48 hours), (b) stage 2 (72 hours), and stage 3 (96 hours) spectral measurements, respectively, based on the difference bands (Figure 5.3 a-c). Each score represents the measured spectra. The achieved classification accuracies of PCA-LDA based on the first twenty selected principal components (PCs) were 97%, 98%, and 98% respectively. The sensitivity measurement values for PC3 cells were 99%, 98%, and 99%, respectively.

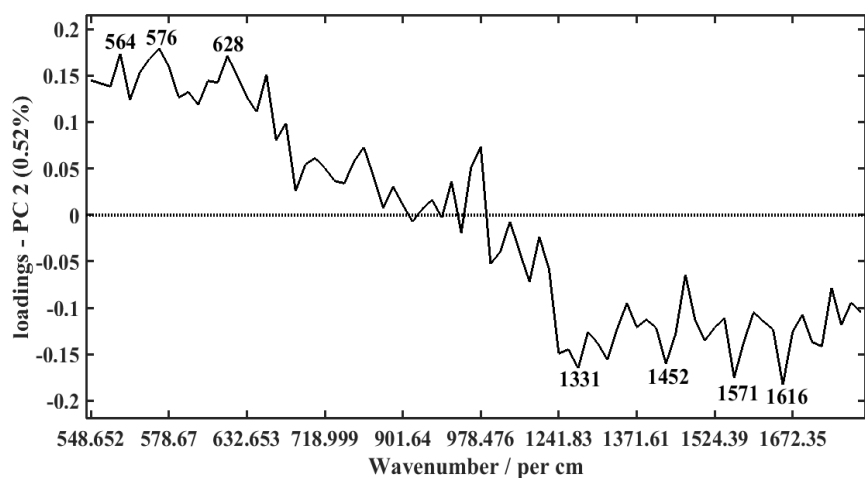
The associated loadings (Figure 5.5(a-c)) shows the bands at (558, 564, 630 cm^{-1}) and (1331, 1337, 1441, 1446, 1449, 1452, 1454 cm^{-1}) dominantly determined the assignment of scores into the malignant (PC3) and normal (PNT1a) classes respectively. The bands at (558-564 cm^{-1} , 630 cm^{-1}) and (1331 - 1337 cm^{-1}) indicated heightened nucleic acid and lipid biochemicals, respectively. The bands within 1440-1460 cm^{-1} region pointed to mixed lipids, nucleic acid and



(a)



(b)



(c)

Figure 5.5 The second principal component (PC 2) loadings that explain discrimination of PC 3 cell scores and PNT1a scores.

protein alterations in both cells. Further, the 1571 cm^{-1} and 1616 cm^{-1} bands strongly determined discrimination of late malignancies. These bands are attributed to (guanine, adenine) and (tyrosine, tryptophan) alterations, respectively (Chandra *et al.*, 2015).

5.1.2 Analysis of subtle biochemical alterations in prostatic cells

As observed in Figure 5.2(b), the 520-715 cm^{-1} , 725-855 cm^{-1} , 860-1000 cm^{-1} , 1085-1250 cm^{-1} , 1255-1445 cm^{-1} , and 1450-1655 cm^{-1} spectral regions were dominated by subtle Raman peaks. A key challenge is singling out weak Raman bands significant for disease diagnostics. Therefore, utility of intermediate- and high-order principal components was adopted as a potential method of mining significant weak variance signals (subtle Raman peaks) for prostate cancer diagnosis. Figure 5.6 shows the eigenvalues of the measured mean Raman spectra plotted as a function of the number of principal components. As highlighted in Section 4.6.2, the prostatic cells datasets were baseline corrected by the linear method and normalized to their maximum intensity, but without denoising to preserve all pertinent spectral features in the data sets. The eigenvalues indicate that about 5 principal components accounted for the largest variance in the spectral datasets. Table 5.3 shows categorization of principal components as either lower, intermediate, or higher-order principal components based on the size of variances and cumulative percentages of the total variation.

Table 5.3 Categorization of principal components (PCs) based on their cumulative percentage of total variation and the variance sizes: low-, intermediate-, and high-order PCs

| Spectral dataset | Low-order PCs (<99% of the cumulative variance and >1.0 average eigenvalue) | Intermediate-order PCs (between 99% and 99.5% of the cumulative variance) | High-order PCs (>99.5% of the cumulative variance and <1.0 average eigenvalue) |
|------------------|---|---|--|
| Stage 1 | 1 | 2 – 16 | 17 – 154 |
| Stage 2 | 1 | 2 – 22 | 23 – 160 |
| Stage 3 | 1 | 2 | 3 – 198 |

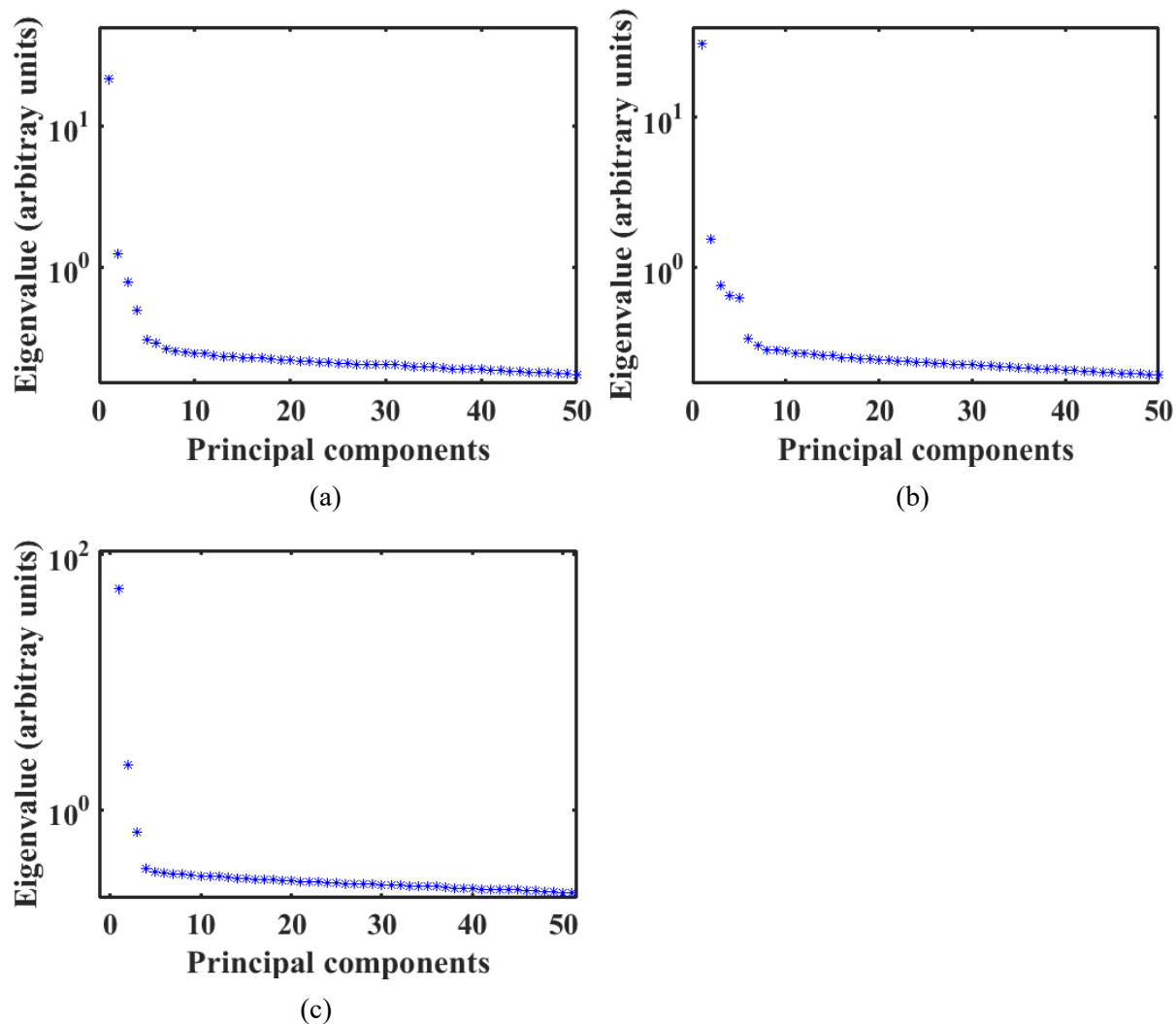


Figure 5.6 Scree plots showing eigenvalues explained as a function of the number of principal components, for (a) stage 1, (b) stage 2, and (3) stage 3 spectral datasets.

It was noted the first principal component (PC 1) could be categorized as a low-order principal component. For stage 1 dataset of both cell lines (PC3 and PNT1a), principal components 2 to 16 were categorized as intermediate - order PCs. For stages 2 and 3 datasets, intermediate - order principal components were 2 to 22, and 2, respectively. The remaining principal components were categorized as higher-order principal components. The first two principal components (PC 1, PC 2) were found to account for the largest variance of the data. Overall, the total cumulative variances accounted for by the first two principal components (PC 1, PC 2) in stage 1, stage 2 and, stage 3 datasets were 99.13%, 99.02%, and 99.58%, respectively, meaning PC 1 and PC 2 explained the prominent biochemical alterations associated with PC3 and PNT1a proliferation. To

determine which PCs scores were potentially significant for sample discrimination, the PCs were subjected to the two sample *t*-test and effect size (Cohen’s *d*, Pearson’s correlation coefficient *r*) statistical criteria, as described in Section 4.6.2. Table 5.4 shows the *p* values and effect sizes computed for the first 10 principal components. Significant differences (*p* <0.05) between the principal component scores were observed in (PCs 1-5), (PCs 1-3, 8), and (PCs 1-6, 8) in stage 1, stage 2 and stage 3 spectral datasets, respectively.

Table 5.4 The Student *t*-test (*p*-values), and effect sizes (Cohen-*d*, Pearson’s correlation coefficients (*r*)) showing the relationship between the principal component scores of normal (PNT1a) and malignant (PC3) cells

| PC | Spectral datasets | | | | | | | | |
|----|--------------------------|---------|----------|--------------------------|---------|----------|--------------------------|---------|----------|
| | Stage 1 | | | Stage 2 | | | Stage 3 | | |
| | <i>p</i> -value | Cohen d | <i>r</i> | <i>p</i> -value | Cohen d | <i>r</i> | <i>p</i> -value | Cohen d | <i>r</i> |
| 1 | 0.003 | 1.22 | 0.52 | 0.04 | 1.32 | 0.55 | 0.04 | 2.27 | 0.75 |
| 2 | 4.04 x 10 ⁻¹⁹ | 0.99 | 0.44 | 2.11 x 10 ⁻³⁹ | 1.18 | 0.51 | 3.56 x 10 ⁻³⁹ | 2.36 | 0.76 |
| 3 | 0.0095 | 0.39 | 0.19 | 9.98 x 10 ⁻¹⁰ | 0.99 | 0.44 | 0.029 | 0.18 | 0.09 |
| 4 | 7.13 x 10 ⁻¹⁸ | 1.17 | 0.50 | 0.42 | 0.03 | 0.02 | 1.51 x 10 ⁻¹⁸ | 1.38 | 0.57 |
| 5 | 3.24 x 10 ⁻¹⁰ | 1.08 | 0.48 | 0.20 | 0.13 | 0.06 | 0.0088 | 0.34 | 0.17 |
| 6 | 0.18 | 0.15 | 0.07 | 0.32 | 0.08 | 0.04 | 0.0082 | 0.19 | 0.09 |
| 7 | 0.21 | 0.13 | 0.06 | 0.36 | 0.06 | 0.03 | 0.413 | 0.03 | 0.02 |
| 8 | 0.09 | 0.21 | 0.11 | 0.03 | 0.29 | 0.14 | 0.02 | 0.03 | 0.01 |
| 9 | 0.20 | 0.14 | 0.07 | 0.23 | 0.41 | 0.20 | 0.46 | 0.02 | 0.01 |
| 10 | 0.11 | 0.19 | 0.09 | 0.33 | 0.07 | 0.04 | 0.44 | 0.02 | 0.01 |

Generally, the first principal component (PC 1) had the large effect sizes in all datasets, which included *d* = 1.22, *d* = 1.32, and *d* = 2.27 for stage 1, 2 and 3 spectral datasets, respectively. Further, the respective explained total variances were 98.85%, 98.88% and 99.19%. The utility of observed PCs in cells discrimination was further evaluated by plotting their canonical variable distributions (Figure 5.7 (a, c, e)).

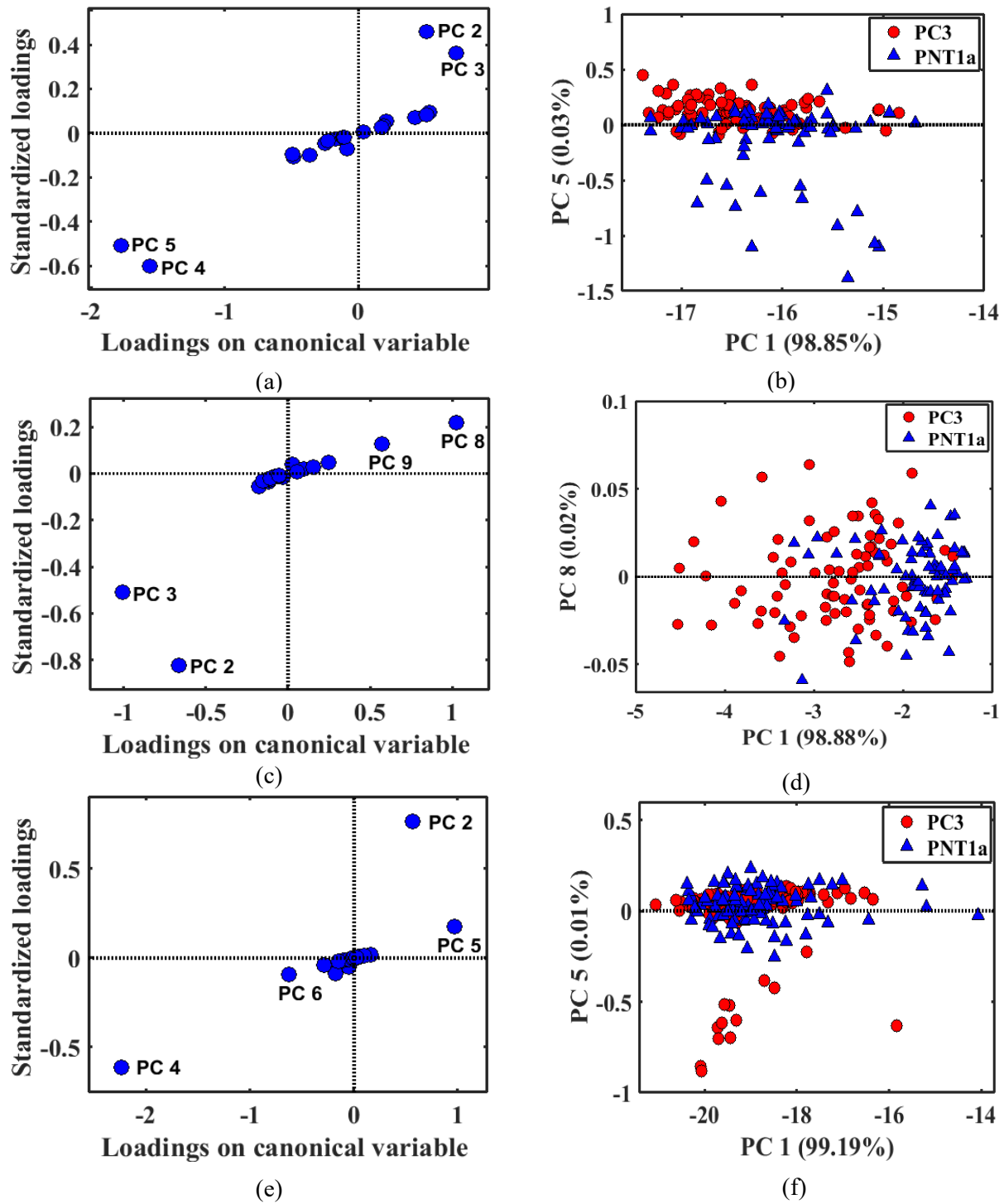


Figure 5.7 The canonical variable distribution plots (a, c, e) showing distinct positions of significant principal components, and the score plots of intermediate and higher-order PCs and the diagnostic line from LDA for (b) stage 1, (d) stage 2, and (f) stage 3 spectral measurements.

The canonical variable distribution plots for stage 1, stage 2, and stage 3 datasets showed PCs (2, 3, 4, 5), (2, 3, 8, 9) and (2, 4, 5, 6) had the largest standardized loadings values, respectively, a parameter that demonstrated their potential strength for cells discrimination according to the level of malignancy. The principal components (2, 3, 4, 5), (2, 3, 8, 9) and (2, 4, 5, 6) in stage 1, 2 and 3 datasets were therefore used for cells discrimination. This was done by examining the scatter plots of PC scores and their respective loadings vectors, where the loading vectors reflects the weights of biochemical components in each spectrum (Corsetti *et al.*, 2018). The score plots highlighted the natural groupings of each of the principal components for the cell types. It was found that intermediate and higher-order PCs 2 (0.29%), 2 (0.42%), and 4 (0.93%) had best grouping of both cells types for stage 1, 2, and 3 spectral data respectively. Examination of loading vectors showed the groupings / discriminations were attributed to the prominent biochemical differences between the cells (Figure 5.3).

It was observed that some level of clustering was present due to PC 5 (0.03%), PC 8 (0.02%) and PC 5 (0.01%) in stage 1, 2, and 3 spectral data respectively (Figure 5.7 b, d, f). These discriminations were attributed to the subtle biochemical differences between the malignant and normal cells. It can be seen that the observed PC scores can be associated with particular loading vectors; where the scores refer to the weight of particular biochemical components in each spectrum. For instance, the few extreme scores defined by Figure 5.7 (d) can be explained by respective prominent bands in the loading vectors spectrum. Given that the aim of this study was to investigate the subtle biochemical alterations (weak variance signals) associated with prostate cancer progression, the following principal components were selected for further analysis: PC 5 (0.03%) for stage 1, PC 8 (0.02%) for stage 2 and PC 5 (0.01%) for stage 3 datasets. The principal component loadings that explained scores discrimination due to PC 5 (0.03%) for stage 1, PC 8 (0.02%) for stage 2 and PC 5 (0.01%) for stage 3 datasets are shown in Figure 5.8 (a), (b), and (c) respectively.

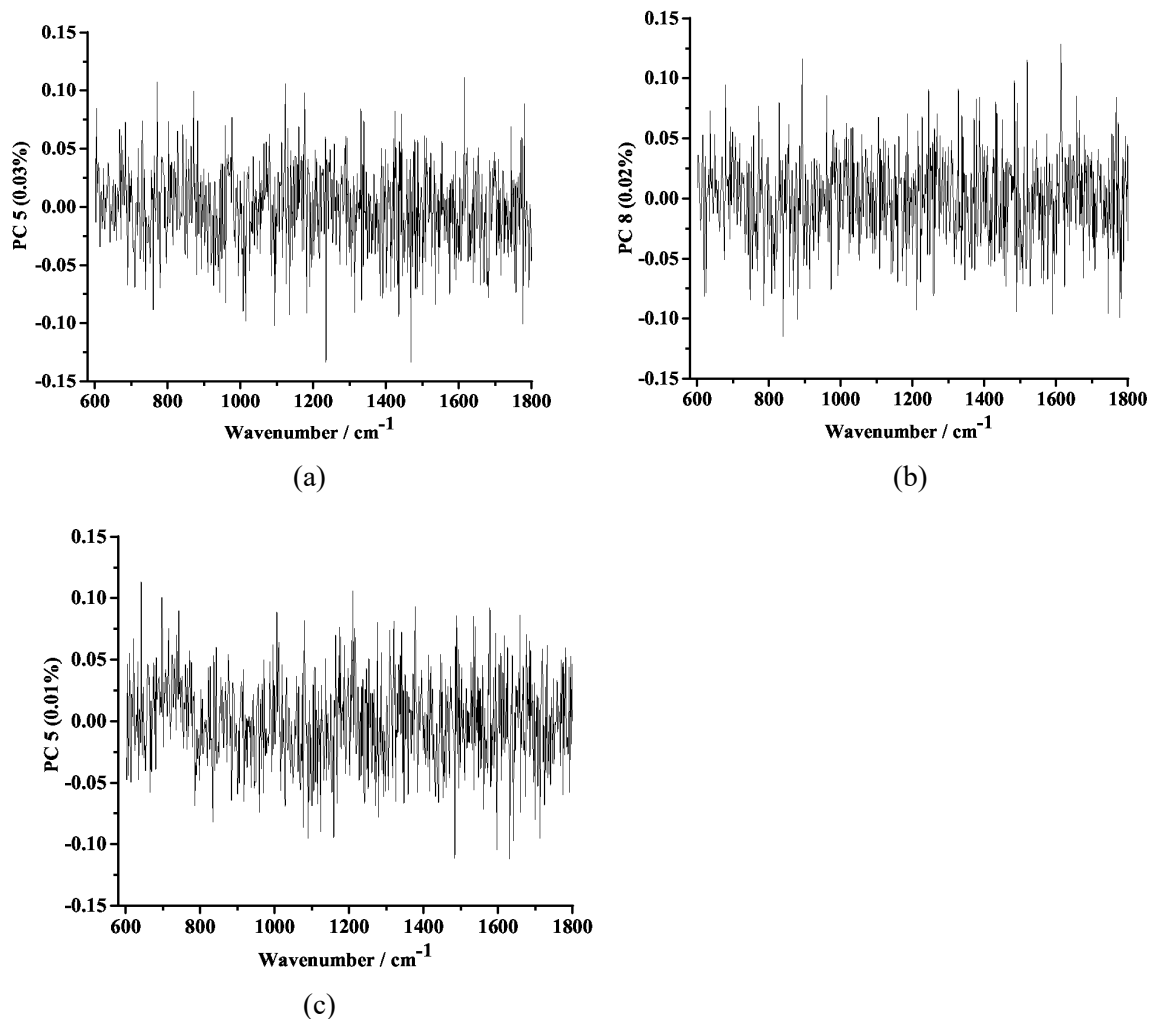


Figure 5.8 Loading vectors explaining the distribution of low- order PC (PC 1) scores against the intermediate-order PCs scores (PC 5 for stage 1, PC 8 for stage 2, PC 5 for stage 3) in (a) stage 1 (48 hours), (b) stage 2 (72 hours) and (c) stage 3 (96 hours) spectral datasets, respectively. Loadings profiles are noted to contain useful peaks amid noisy features. The useful peaks were extracted by adjustment of threshold and sensitivity levels / parameters in *peak finding* function in Omnic[®] software from Thermo Scientific. Sensitivity levels were set at a low value (30%) so that noise and other unimportant features above threshold value were eliminated.

As previously explained, the Cohen d effect sizes were classified as small ($d=0.2$), medium ($d = 0.5$), and large ($d \geq 0.8$) (Sullivan *et al.*, 2012). Thus, Table 5.4 shows PC 5 (0.03%) for stage 1 dataset had a large effect size whereas PC 8 (0.02%) for stage 2 and PC 5 (0.01%) for stage 3 datasets had medium effect sizes. The loadings profiles (Figure 5.8 (a-c)) were noted to contain useful peaks amid noisy features. A possible method of extracting useful peaks from noise would be to denoise or smooth the loading profiles. However, this would introduce spectral artifacts, peak shifts, and loss of useful peaks. Therefore, the peaks were extracted in their raw form by adjusting threshold and sensitivity levels / parameters of *peak finding* function in Omnic[®] software from Thermo Scientific. It should be noted the sensitivity levels were kept at low values (30%) to eliminate noisy peaks, and only the most intense loading vectors that explained scores distribution in Figure 5.7 (b, d, f) were extracted and used for further analysis. Detailed information concerning the loading vectors (which in this case are weak Raman bands) that explained natural groupings of PNT1a and PC3 scores using intermediate-and high-order PC 5 (0.03%) for stage 1, PC 8 (0.02%) for stage 2 and PC 5 (0.01%) for stage 3 datasets are shown in Table 5.5.

It was observed (Table 5.5) that the loading vectors (Raman bands) at 771 cm^{-1} , 871 cm^{-1} , 1122 cm^{-1} , 1176 cm^{-1} , 1614 cm^{-1} (for stage 1), 1068 cm^{-1} , 1191 cm^{-1} , 1333 cm^{-1} , 1471 cm^{-1} , 1524 cm^{-1} , 1586 cm^{-1} (for stage 2) and 1076, 1089 cm^{-1} , 1278 cm^{-1} , 1596 cm^{-1} , 1631 cm^{-1} , 1712 cm^{-1} (for stage 3) had the most influence for the assignment of scores into the malignant class (PC3 cells). Likewise, the loading vectors at (1014, 1092, 1232, 1468, 1776 cm^{-1}), (704, 1217, 1271, 1274, 1531 cm^{-1}) and (641, 697, 1319, 1377, 1577 cm^{-1}) had the most influence for the assignment of scores into the normal class (PNT1a) for stage 1, 2, and 3 measurements respectively. It should be noted that the loading vectors at (1232, 1330, 1442 cm^{-1}), (1471 cm^{-1}), and (1076, 1278 cm^{-1}) were common for the two cell lines in stage 1, 2, and 3 datasets respectively. The two sample t -test on averaged normalized intensities on these bands indicated they were statistically significant (Students t -test, $p < 0.05$).

To better understand these differences, band intensity ratios (I_C / I_N) were determined at all observed loading vectors (weak Raman bands), by dividing the normalized intensities of diseased (PC3 cells) Raman spectra (I_C) by the respective normalized intensities of control (PNT1a cells) Raman spectra (I_N). The band intensity ratio values at 1076 cm^{-1} and 1232 cm^{-1} bands were found to increase with the stage of cell proliferation (Table 5.6).

Table 5.5 The Raman bands (loading vectors) that explain natural groupings of PNT1a and PC3 scores using the fifth principal component (PC 5 (0.03%)) for stage 1, eighth principal component (PC 8 (0.02%)) for stage 2 and fifth principal component (PC 5 (0.01%)) for stage 3 spectral datasets (the values are in units of cm^{-1})

| Prostatic cells | | | | | |
|-------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| PC3 | | | PNT1a | | |
| Stage 1 | Stage 2 | Stage 3 | Stage 1 | Stage 2 | Stage 3 |
| 604 | 508 | 574 | 1007 | 517 | 641* |
| 668 | 564 | 786 | 1014* | 704** | 697* |
| 730 | 869 | 833 | 1092* | 716 | 1076 ^c |
| 771* | 913 | 884 | 1232 ^a | 752 | 1209 |
| 841 | 1068* | 900 | 1330 ^a | 772 | 1278 ^c |
| 871* | 1077 | 969 | 1391 | 1217* | 1319* |
| 976 | 1191* | 1027 | 1442 ^a | 1271* | 1377* |
| 1122* | 1333* | 1076 ^c | 1468* | 1274* | 1577* |
| 1176* | 1471 ^b | 1089* | 1681 | 1364 | |
| 1232 ^a | 1524* | 1094 | 1776* | 1471 ^b | |
| 1330 ^a | 1586* | 1100 | | 1514 | |
| 1339 | | 1123 | | 1531* | |
| 1442 ^a | | 1158 | | 1557 | |
| 1614* | | 1240 | | | |
| | | 1269 | | | |
| | | 1278 ^c | | | |
| | | 1439 | | | |
| | | 1482 | | | |
| | | 1561 | | | |
| | | 1596* | | | |
| | | 1631* | | | |
| | | 1641 | | | |
| | | 1699 | | | |
| | | 1712* | | | |
| | | 1724 | | | |

KEY: The asterisks (*) identify the loading vectors (weak Raman bands) that had the most influence in the assignment of scores for each of the two cell lines. Superscripts a, b, c identify the loading vectors (weak Raman bands) that were common for both cell lines.

Table 5.6 Comparison of peak intensity ratios between malignant (PC3) and normal cells (PNT1a) based on the subtle Raman peaks at 1076 cm⁻¹ and 1232 cm⁻¹. The numbers 1, 2, and 3 enclosed in brackets identify the stage of cell proliferation

| | Stage 1 | | Stage 2 | | Stage 3 | | | | |
|------------------------------------|-------------------------------------|------|---------|------|---------|------|---------------------------|---------------------------|---------------------------|
| | Raman intensities (arbitrary units) | | | | | | Band ratios* | | |
| Raman shift (cm ⁻¹) | PNT1a | PC3 | PNT1a | PC3 | PNT1a | PC3 | $\frac{PC3(1)}{PNT1a(1)}$ | $\frac{PC3(2)}{PNT1a(2)}$ | $\frac{PC3(3)}{PNT1a(3)}$ |
| 1076 | 0.67 | 0.65 | 0.64 | 0.64 | 0.74 | 0.77 | 0.970 | 1.0 | 1.04 |
| 1232 | 0.60 | 0.55 | 0.74 | 0.68 | 0.79 | 0.78 | 0.917 | 0.919 | 0.987 |

KEY: The asterisks (*) identify the band ratios computed by dividing the normalized intensities of diseased (PC3 cells) Raman spectra (I_C) by the respective normalized intensities of control (PNT1a cells) Raman spectra (I_N).

The 1076 cm⁻¹ band is linked to stretching modes of phosphate components that emanate from nucleic acids and is thought to suggest nucleic acid levels heighten with malignancy (Movasaghi *et al.*, 2007). The alterations around 1232 cm⁻¹ band indicate antisymmetric phosphate stretching vibrations due to nucleic acids and Amide III alterations (Rehman *et al.*, 2013). Therefore, the increasing ratios of normalized intensities of PC3 cells (I_C) to normalized intensities of PNT1a cells (I_N) suggest prostate malignancy can be associated with an increase in relative amounts of nucleic acids and Amide III alterations. The rest of loading vectors (Table 5.5) can be mainly attributed to subtle nucleic acids and protein alterations, and their specific assignments are explained elsewhere (Chandra *et al.*, 2015; Movasaghi *et al.*, 2007; Gelder *et al.*, 2007).

The use of high-order principal components is still underexplored as a machine learning tool for the analysis of prostatic tissues. In fact, prostate cancer Raman studies prefer use of the first few principal components (PCs) because they account for large proportions of variance (Theophilou *et al.*, 2015; Musto *et al.*, 2017; Corsetti *et al.*, 2018; Matias *et al.*, 2011; Crow *et al.*, 2005; Patel *et al.*, 2010). However, adding a supervised constraint on the PCA e.g., LDA, for discrimination purpose increases weight of underlying spectral features in the classification by eliminating variations in spectral features not correlated to a pathological state but other less relevant biological parameters (Nargis *et al.*, 2019). A previous review on applications of Raman spectroscopy for prostate cancer (Kast *et al.*, 2014), shows the majority of prostate-based Raman

studies reported dominance of proteins, lipids, and nucleic acids alterations, especially at 1000 cm^{-1} , 1200-1350 cm^{-1} , 1450 cm^{-1} , and 1600-1700 cm^{-1} regions. To the best of authors' knowledge, spectroscopic prostate studies based on the utility of higher-order PCs; which account for negligible variance (assumed noise) have not been reported. In this study, the observed subtle bands can be explained in consideration of different experimental conditions that could have limited detection of optimum Raman signals e.g., scattering efficiency of laser wavelength (thus the laser power irradiation over the sample), and the large background due to scattering from the sample itself, substrate and the microscope objective (Byrne *et al.*, 2015). Generally, the large background effects may swamp any weak Raman signals present, although they could be of diagnostic value. Further, useful Raman bands may not be optimally detected during experiments.

For instance, previous closely related work on prostate cancer cells by Corsetti *et al.*, (2018) and Crow *et al.*, (2005) reported related prominent bands around (880, 1184, 1588, 1614 cm^{-1}) and (1094 / 96, 1125, 1576 cm^{-1}) respectively. Similarly, the 1328 cm^{-1} and 643 cm^{-1} bands were observed as prominent bands by Matias *et al.*, (2011) and Patel *et al.*, (2010), respectively. Although the present study detected similarly closely related loading vectors which include 871 cm^{-1} , 1094 cm^{-1} , 1122 / 1123 cm^{-1} , 1191 cm^{-1} , 1330 / 1333 cm^{-1} , 1586 cm^{-1} and 1614 cm^{-1} as shown in table 5.5, their Raman intensities were generally subtle, therefore could not be detected in spectral profiles shown in Figure 5.2 (b) and Figure 5.3 (a-c). However, it should be noted the works by Corsetti *et al.*, (2018), Crow *et al.*, (2005), Matias *et al.*, (2011) and Patel *et al.*, (2010) were based on 785 nm, 830 nm, 830 nm and 785nm excitation lasers of 135 mW, 300 mW, 80 mW and 100 mW laser powers, respectively. The intensity of Raman scattered radiation is directly proportional to the frequency of incident radiation (equation 3.17). Therefore, it is not clear if the excitation energies at the point of sample surface; and which have not been explicitly stated by the authors (Corsetti *et al.*, 2018; Crow *et al.*, 2005; Matias *et al.*, 2011; Patel *et al.*, 2010), played a major role in detection of optimum Raman signals in their respective findings. Our results showed the loading vectors in Table 5.5 were not spectrally visible in Figure 5.2 b and Figure 5.3 a-c, although were statistically significant ($p < 0.05$). However, as observed (Figure 5.7 (b, d, f)), the score plots demonstrated a reasonable level of scores discrimination; significantly strengthening the view that there were inherent subtle molecular differences between the two cell lines. These results were encouraging and motivated application of intermediate- and higher-order principal components in extracting useful weak band variance signals (weak spectral markers) in whole blood and saliva spectral datasets for breast cancer and leukemia diagnostics.

5.2 Raman spectroscopic characterization of blood and saliva fluids for breast cancer diagnostics

5.2.1 Raman spectroscopy characterization of whole blood

5.2.1.1 Analysis of prominent biochemical alterations in whole blood spectra

The optical photomicrograph of a typical blood sample is shown (Figure 5.9 (a)). The characteristic pinkish red color of blood smear is observed, a factor attributed to hemoglobin due to iron minerals. Figure 5.9 (b) shows stacked averaged Raman spectrum of control and diseased samples in 400 – 1800 cm^{-1} region, with spectral data normalized at 1446 cm^{-1} band (lipids / proteins). The biochemical assignments of peaks in Figure 5.9 (b-d) were done in accordance with the Raman spectroscopy of tissues, body fluids, or bio-molecules as highlighted in literature (Pichardo-Molina *et al.*, 2007; Vargas-Obieta *et al.*, 2016; Rehman *et al.*, 2013; Gelder *et al.*, 2007). Detailed information regarding Raman band assignments are provided in Table 5.7. As seen in Figure 5.9 (b), the control samples spectra exhibited higher intensity peaks attributable to nucleic acids (744, 1339, 1574 cm^{-1}), phospholipids (744, 965, 1124, 1339, 1446 cm^{-1}), proteins (1124, 1240, 1339, 1446, 1574 cm^{-1}) and saccharides (410, 479, 1124 cm^{-1}) when compared to diseased samples spectra. On the other hand, diseased samples spectra had higher intense bands attributed to proteins (1002, 1617 cm^{-1}), phospholipids (1367 cm^{-1}) when compared to control samples spectra.

To identify specific constituents explaining biochemical changes in blood of control and diseased patients, the spectra differences between the control and diseased samples in the 500-1800 cm^{-1} region were considered (Figure 5.9 (c)). Based on literature (Movasaghi *et al.*, 2007; Chandra *et al.*, 2015; Gelder *et al.*, 2007), it was observed the bands attributable to saccharides (408, 479 \pm 0.47, 1125 \pm 2.05 cm^{-1}), nucleic acids (742 \pm 2.27, 1341 \pm 1.08, 1583 \pm 2.24 cm^{-1}), phospholipids (742 \pm 2.27, 968, 1125 \pm 2.05, 1341 \pm 1.08 cm^{-1}), proteins (998 \pm 0.70, 1125 \pm 2.05, 1244 \pm 0.70, 1341 \pm 1.08, 1583 \pm 2.24 cm^{-1}) explained heightened biochemical alterations in control samples, validating observations in Figure 5.9(b). Besides, the strongest band occurred at 1244 \pm 0.70 cm^{-1} and 1583 \pm 2.24 cm^{-1} which suggested that biochemical changes due to proteins (amide III, tyrosine, arginine) were predominant in control and diseased samples. The diseased samples were found to exhibit heightened band at 845 \pm 1.31 cm^{-1} assigned to tyrosine proteins.

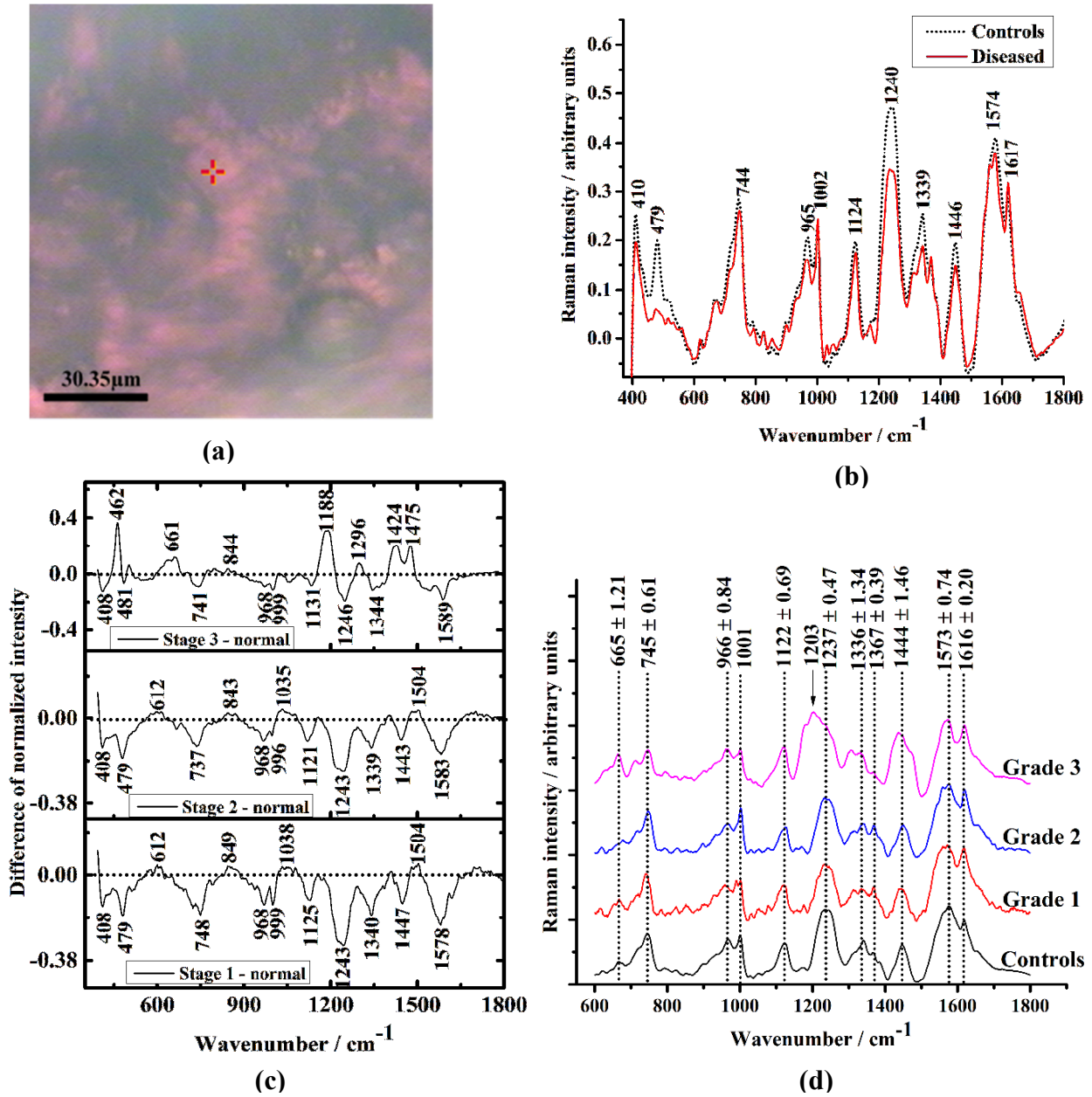


Figure 5.9 (a) The optical photomicrograph of a dried blood sample with a laser spot (+) indicated (50x magnification), (b) the overall stacked mean spectra (normal: ($n = 23$); diseased: ($n = 20$), (c) Raman alterations in controls ($n = 23$), grade 1 ($n = 3$), grade 2 ($n = 7$), and grade 3 ($n = 10$) diseased samples, and (d) spectra differences between Raman spectra of healthy and diseased samples.

The blood spectrum \pm standard deviations (SD) of controls ($n = 23$), grade 1 ($n = 3$), grade 2 ($n = 5$), and grade 3 ($n = 10$) groups were compared, in fingerprint region - $500 - 1800\text{cm}^{-1}$ (Figure 5.9 (d)). Note that spectra have been linearly offset for clarity. There were 11 primary bands ($p < 0.05$) at the wavelengths around $665 \pm 1.21 \text{ cm}^{-1}$ (nucleic acids, phospholipids), $745 \pm 0.61 \text{ cm}^{-1}$ (nucleic acids, phospholipids), $996 \pm 0.84 \text{ cm}^{-1}$ (C-O ribose, C-C stretch), 1001 cm^{-1} (phenylalanine), $1122 \pm 0.69 \text{ cm}^{-1}$ (proteins, lipids, glucose), $1237 \pm 0.47 \text{ cm}^{-1}$ (proteins), $1336 \pm 1.34 \text{ cm}^{-1}$ (proteins, phospholipids, nucleic acids), $1367 \pm 0.39 \text{ cm}^{-1}$ (phospholipids), $1444 \pm 1.46 \text{ cm}^{-1}$ (phospholipids, proteins), $1573 \pm 0.74 \text{ cm}^{-1}$ (nucleic acids, proteins) and $1616 \pm 0.20 \text{ cm}^{-1}$ (proteins). Besides, there was a spectral feature at 1203 cm^{-1} uniquely in grade 3 breast cancer profile, which can be majorly attributed to protein alterations. The protein bands were based on aromatic acids (e.g. tryptophan, phenylalanine, and tyrosine), while nucleic acid alterations were mainly due to DNA and RNA bases (cytosine, uracil, adenine, thymine, guanine).

As noted in Figure 5.9(d), enhanced peaks were predominantly detected in $600-1800\text{cm}^{-1}$ region and the strongest bands occurred at $745 \pm 0.61 \text{ cm}^{-1}$, 1001 cm^{-1} , $1237 \pm 0.47 \text{ cm}^{-1}$, $1444 \pm 1.46 \text{ cm}^{-1}$, $1573 \pm 0.74 \text{ cm}^{-1}$, and $1616 \pm 0.20 \text{ cm}^{-1}$. Our finding correspond to (Pichardo-Molina *et al.*, 2007; Nargis *et al.*, 2019; Vargas-Obieta *et al.*, 2016) where prominent peaks in blood and serum spectra of healthy volunteer controls and patients clinically diagnosed with breast cancer were majorly observed in $600-1800 \text{ cm}^{-1}$ region, and bands around 1001 cm^{-1} and 1444 cm^{-1} were reportedly heightened in all spectra profiles. But, contrary to previous findings (Pichardo-Molina *et al.*, 2007; Nargis *et al.*, 2019; Bilal *et al.*, 2017), the 1658 cm^{-1} band was not detected in our work, a disparity we attributed to the potential compositional differences between whole blood and serum samples and the oxygenation state of hemoglobin in whole blood samples (Atkins *et al.*, 2017).

Table 5.7 Raman band assignments of healthy people and breast cancer patients

| Raman shift (cm^{-1}) | Functional groups and molecular vibration assignments | References |
|--|--|---|
| 408, 410 | Saccharides | (Gelder <i>et al.</i> , 2007) |
| 462 | Ring breathing modes of phenylalanine | (Rehman <i>et al.</i> , 2013) |
| 479, 483, 479 \pm 0.47 | $\nu(\text{COH})$, $\nu(\text{CCH})$, $\nu(\text{OCH})$ side group deformations of saccharides | (Rehman <i>et al.</i> , 2013) (Gelder <i>et al.</i> , 2007) |
| 612 | Cholesterol esters | (Rehman <i>et al.</i> , 2013) |
| 661, 665 \pm 1.21 | C-S stretching modes (proteins), Phospholipids, ring breathing modes of DNA / RNA bases (thymine, guanine) | (Pichardo-Molina <i>et al.</i> , 2007) (Rehman <i>et al.</i> , 2013) |
| 744, 742 \pm 2.27, 745 \pm 0.61 | C-S stretch of phospholipids, Ring breathing modes of DNA bases | (Pichardo-Molina <i>et al.</i> , 2007) (Rehman <i>et al.</i> , 2013) |
| 845 \pm 1.31 | Single-bond stretching vibrations for tyrosine | (Pichardo-Molina <i>et al.</i> , 2007) |
| 965, 968 | $\delta(=\text{CH}$ wagging) of lipids | (Rehman <i>et al.</i> , 2013) |
| 996 \pm 0.84 | C-O stretch of ribose, C-C stretch | (Rehman <i>et al.</i> , 2013) |
| 998 \pm 0.70, 1002 | Symmetric ring breathing mode of phenylalanine | (Vargas-Obieta <i>et al.</i> , 2016), |
| 1036 \pm 0.86 | CH_2CH_3 bending modes of phenylalanine and collagen | (Pichardo-Molina <i>et al.</i> , 2007) |
| 1122 \pm 0.69, 1125 \pm 2.05 | C-N stretch of proteins, C-C stretch of lipids and proteins, Glucose | (Vargas-Obieta <i>et al.</i> , 2016), (Rehman <i>et al.</i> , 2013) |
| 1188 | Antisymmetric phosphate vibrations of DNA and RNA bases cytosine, guanine, adenine) | (Rehman <i>et al.</i> , 2013) |
| 1203 | Amide III, CH_2 wagging (glycine, proline) Tyrosine, Phenylalanine | (Rehman <i>et al.</i> , 2013) |
| 1237 \pm 0.47, 1240, | β – sheet. Amide III, CH_2 deformations of glycine and proline | (Vargas-Obieta <i>et al.</i> , 2016), |
| 1244 \pm 0.70 | β – sheet. Amide III, CH_2 deformations of glycine and proline | (Pichardo-Molina <i>et al.</i> , 2007) |

Table 5.7 (continued) Raman band assignments of healthy people and breast cancer patients

| Raman shift (cm^{-1}) | Functional groups and molecular vibration assignments | References |
|---|---|---|
| 1296 | CH ₂ deformation of tryptophan, α – helix, and phospholipids | (Vargas-Obieta <i>et al.</i> , 2016) |
| 1336 \pm 1.34, 1339, 1341 \pm 1.08 | δ (CH ₃), δ (CH ₃), twisting of proteins and phospholipids, α – helix, adenine, guanine | (Vargas-Obieta <i>et al.</i> , 2016) (Rehman <i>et al.</i> , 2013) |
| 1367 \pm 0.39 | ν_s (CH ₃) stretch of phospholipids | (Rehman <i>et al.</i> , 2013) |
| 1424 | CH ₂ bending mode of proteins and lipids | (Rehman <i>et al.</i> , 2013) |
| 1444 \pm 1.46, 1446 | δ (CH ₂ / CH ₃), scissoring ((phospholipids), and proteins | (Pichardo-Molina <i>et al.</i> , 2007) |
| 1504 | C-C stretch of phenylalanine, ring breathing modes of cytosine | (Rehman <i>et al.</i> , 2013) |
| 1573 \pm 0.74 | Ring breathing modes of DNA / RNA bases (guanine, adenine), C=C bending modes of tryptophan protein | (Rehman <i>et al.</i> , 2013) |
| 1583 \pm 2.24 | C=C bending modes of phenylalanine tyrosine, arginine, and adenine | (Pichardo-Molina <i>et al.</i> , 2007) |
| 1616 \pm 0.20, 1617 | C = C stretching of tyrosine and tryptophan | (Vargas-Obieta <i>et al.</i> , 2016) (Rehman <i>et al.</i> , 2013) |
| 1754 | C=O stretch of lipids | (Rehman <i>et al.</i> , 2013) |

If we consider grade 1 and grade 2 breast cancer spectra, the diseased samples had common intense bands around 612 cm^{-1} (cholesterol esters), 1036 \pm 0.86 cm^{-1} (phenylalanine proteins), and 1504 cm^{-1} (proteins, nucleic acids). In contrast, the late malignancy (grade 3 breast cancer) exhibited heightened bands around 462 cm^{-1} (phenylalanine), 661 cm^{-1} (nucleic acids, phospholipids), 1188 cm^{-1} (nucleic acids), 1296 cm^{-1} (proteins, phospholipids), 1424 cm^{-1} (proteins, lipids) and 1475 cm^{-1} (δ CH₂ modes), with the strongest bands occurring at 462 cm^{-1} and 1188 cm^{-1} . The disparity between heightened alterations for diseased samples in grade 1, grade 2, and grade 3 breast cancer spectra suggests presence of other biochemical differences during late progression of breast malignancy.

As observed in Figure 5.9(d), there were several notable prominent spectral markers. However, there were spectral regions with weak Raman peaks, for instance, the 750-960 cm^{-1} and 1010-1110 cm^{-1} regions. Therefore, to investigate the possible significant weak (subtle) biochemical alterations useful for scores discrimination, we employed the potential of intermediate and higher-order principal components for disease diagnostics as described in Section 4.6.2.

5.2.1.2 Analysis of trace biochemical alterations in whole blood spectra

The principal component analysis of spectral datasets was performed in 500-1800 cm^{-1} region. The log scree plots that explained scores in spectral datasets for blood samples in grade₁, grade₂ and grade₃ spectral datasets are shown in Figure 5.10.

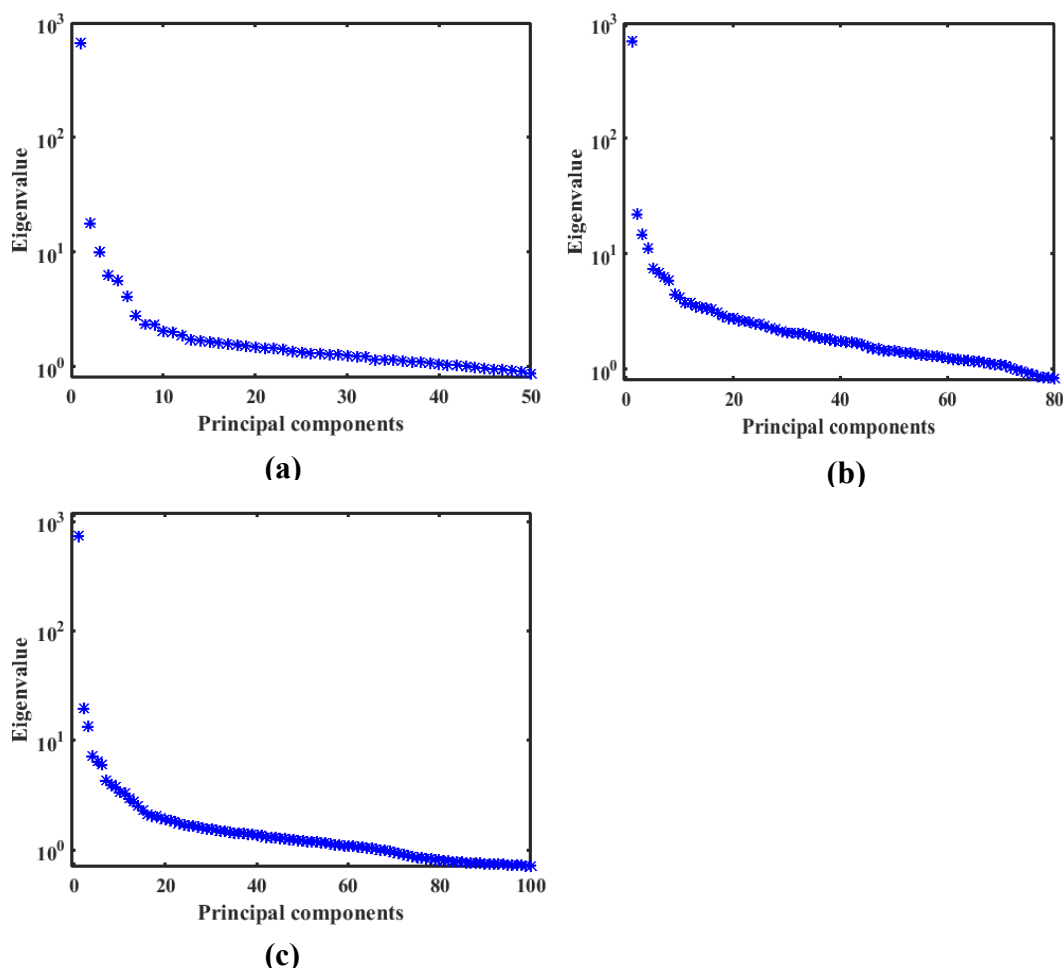


Figure 5.10 The log scree plots that explain scores in (a) grade₁, (b) grade₂ and (c) grade₃ spectral datasets for blood samples from healthy (normal) and breast cancer patients. It was noted that the number of principal components (PCs) with eigenvalue >1 were 42, 72, and 66, respectively.

The determined principal components were categorized as lower-, intermediate-, and higher-order principal components based on the size of variances and cumulative percentages of the total variation criteria as shown in Table 5.8. Moreover, variance method was also applied (Martinez *et al.*, 2005), where it was noted that 12, 24 and 17 PCs could be retained for further analysis of stage₁, stage₂, and stage₃ spectral datasets, respectively. The strengths and directions of correlations between the principal components were examined by use of canonical variable distribution plots, shown in Figure 5.11.

Table 5.8 Categorization of PCs based on the cumulative percentage of total variation and the size of variances: Low-order PCs (<90% of the cumulative variance and >1.0 average eigenvalue), Intermediate-order PCs (between 90% and 95% of the cumulative variance), Higher-order PCs (>95% of the cumulative variance)

| Spectral dataset | low-order PCs | Intermediate-order PCs | Higher-order PCs |
|--------------------|---------------|------------------------|------------------|
| Grade ₁ | 1-2 | 3-18 | 19-300 |
| Grade ₂ | 1-5 | 6-15 | 16-310 |
| Grade ₃ | 1-4 | 5-15 | 16-324 |

It is observed the closely clustered principal components were strongly correlated, and their significance was found to increase with their distinct positions on the plane. The determined statistical values i.e. *t*-test (*p*-values), effect sizes (Cohen-*d*), and Pearson’s correlation coefficients (*r*) that shows relationship between the principal component scores of healthy and diseased blood samples are shown in Table 5.9. The first two principal components (PC1, PC2) explained large fractions of the data; accounting total cumulative variances of 85.51%, 82.72%, 84.27% for grade₁, grade₂, and grade₃ spectral datasets, respectively. Significant differences (*p* < 0.05) between healthy and diseased samples scores were mainly observed in the lower and intermediate order PCs as shown in Table 5.9. As observed, the PCs 3, 2, and 3 were the most significant for grade₁, grade₂, and grade₃ spectral datasets, respectively. Also, they had the largest effect sizes and represented 4.51%, 9.84%, and 3.10% of the total variance in their respective input data. Further, these PCs had intense canonical loading parameters, which confirms their potential strength in samples discrimination (Dattalo, 2014).

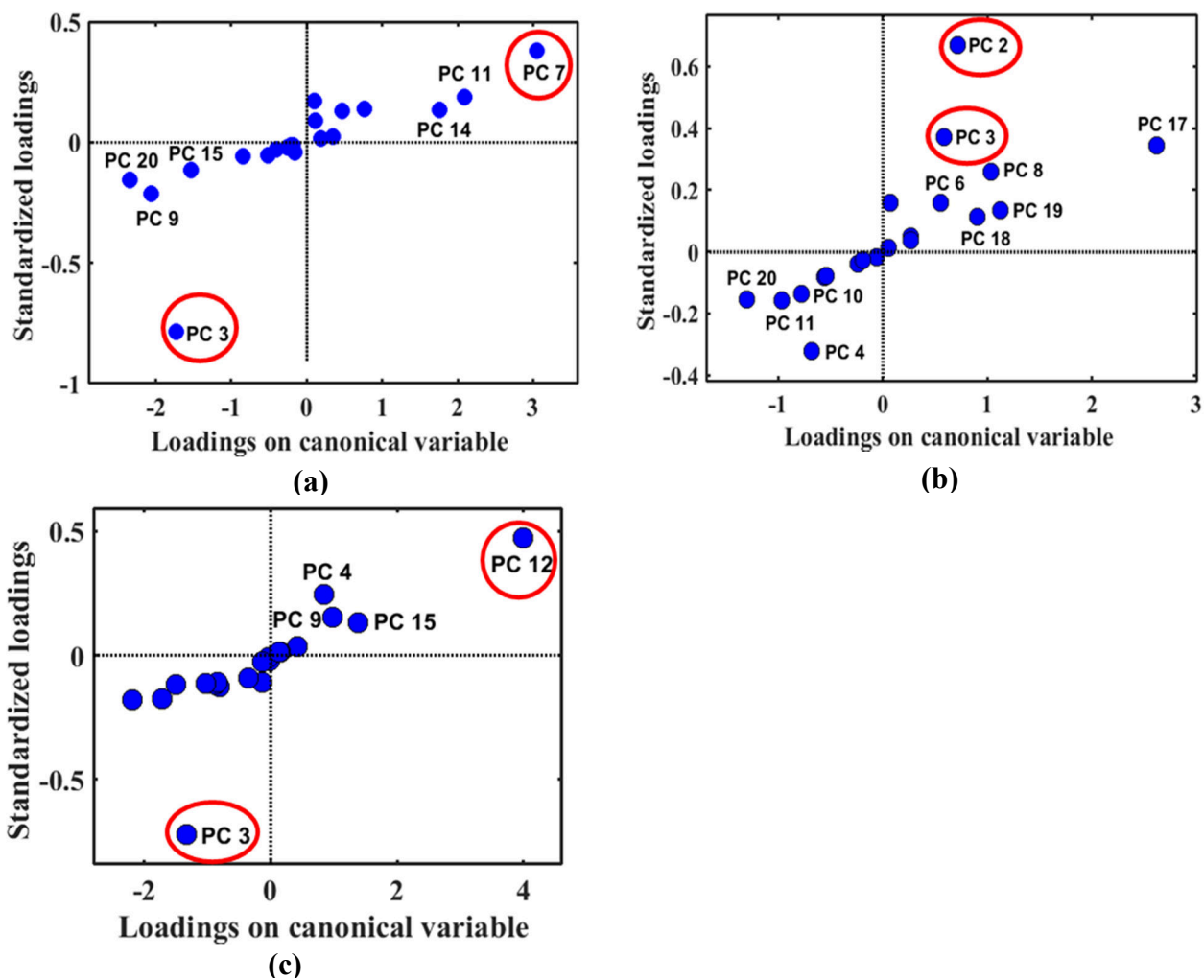


Figure 5.11 Canonical variable distribution showing the low-, intermediate- and high-order principal components for (a) grade₁, (b) grade₂, and (c) grade₃ spectral datasets, respectively. It is observed PCs 3, 2, and 3 have largest canonical loading parameters, suggesting their potential strength for higher classification accuracies in samples discrimination which is validated by their higher levels of statistical significance and effect sizes (Table 5.9). PCs 7, 3, and 12 have fairly higher levels canonical loading parameters and statistical significances, therefore potentially useful for discriminating scores due to presence of subtle biochemical alterations (weak variance signals).

Table 5.9 The *t*-test (*p*-values), and effect sizes ((Cohen-*d*, Pearson’s correlation coefficients (*r*)) showing relationship between the principal component scores of control and diseased blood samples. For clarity, only the statistical values for statistically significant PCs (*p* < 0.05) are shown

| PC | Grade 1 | | | Grade 2 | | | Grade 3 | | |
|----|--------------------------|-----------------|----------|--------------------------|-----------------|----------|--------------------------|-----------------|----------|
| | <i>p</i> - value | Cohen- <i>d</i> | <i>r</i> | <i>p</i> - value | Cohen- <i>d</i> | <i>r</i> | <i>p</i> - value | Cohen- <i>d</i> | <i>r</i> |
| 1 | 6.07 x 10 ⁻⁶ | 0.63 | 0.30 | 0.02 | 0.21 | 0.10 | 0.0085 | 0.21 | 0.10 |
| 2 | 9.59 x 10 ⁻⁷ | 0.69 | 0.32 | 1.27 x 10 ⁻³⁶ | 1.45 | 0.58 | 0.02 | 0.17 | 0.08 |
| 3 | 3.25 x 10 ⁻⁸⁴ | 3.41 | 0.86 | 8.91 x 10 ⁻²² | 1.06 | 0.46 | 1.09 x 10 ⁻⁶⁹ | 1.84 | 0.67 |
| 4 | <i>p</i> > 0.05 | | | 5.96 x 10 ⁻¹⁹ | 0.97 | 0.43 | 8.68 x 10 ⁻⁸ | 0.48 | 0.23 |
| 5 | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | | 0.01 | 0.19 | 0.09 |
| 6 | 0.01 | 0.31 | 0.15 | 4.93 x 10 ⁻⁵ | 0.42 | 0.20 | <i>p</i> > 0.05 | | |
| 7 | 6.94 x 10 ⁻¹⁷ | 1.22 | 0.52 | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | |
| 8 | <i>p</i> > 0.05 | | | 2.1 x 10 ⁻¹⁰ | 0.68 | 0.32 | 0.002 | 0.25 | 0.12 |
| 9 | 4.09 x 10 ⁻⁶ | 0.64 | 0.30 | <i>p</i> > 0.05 | | | 0.0005 | 0.30 | 0.14 |
| 10 | <i>p</i> > 0.05 | | | 0.01 | 0.24 | 0.12 | 0.004 | 0.23 | 0.11 |
| 11 | 1.65 x 10 ⁻⁵ | 0.60 | 0.28 | 0.00 | 0.31 | 0.15 | <i>p</i> > 0.05 | | |
| 12 | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | | 5.3 x 10 ⁻²⁷ | 1.02 | 0.45 |
| 13 | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | | 0.006 | 0.22 | 0.11 |
| 14 | 0.000929 | 0.44 | 0.21 | 0.02 | 0.21 | 0.10 | 7.59 x 10 ⁻⁵ | 0.34 | 0.17 |
| 15 | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | | 0.002 | 0.25 | 0.12 |
| 16 | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | |
| 17 | <i>p</i> > 0.05 | | | 3.98 x 10 ⁻¹² | 0.85 | 0.39 | 5.85 x 10 ⁻⁵ | 0.54 | 0.26 |

The LDA was applied to the PCA results, i.e., to low and intermediate order PCs scores as a technique of supervision. This ensured that the distribution of the data across the scatter plot was due to variations in spectral features correlated to a pathological state but not due to other less relevant biological parameters (Nargis *et al.*, 2019). The linear discriminant analysis distinguished the healthy and diseased groups as shown (Figure 5.12 a-f). Analysis of official reports from pathologists showed the point group marked by blue dots could be associated with control patients’ spectra and the group marked by red triangles could be associated with breast cancer patients’ spectra. The overall classification accuracies for grade₁, grade₂ and grade₃ spectral datasets were

100%, 97% and 93%, respectively, whereas sensitivity and specificity values were (98%, 100%), (84%, 100%), and (82%, 98%), respectively. The best LDA discrimination was obtained by plotting scores of the first principal component (PC 1) versus PC 3 (4.51%), PC 2 (9.84%), and PC 3 (3.10%) scores, for grade₁, grade₂ and grade₃ spectral datasets, respectively, as shown in Figure 5.12 a, c, e. The loading plots were analyzed to understand the prominent and subtle biochemical alterations contributing to these score discrimination. The loading vectors, which explain the weights of biochemical components in spectrum (Corsetti *et al.*, 2018), are shown in Figure 5.13 (a-c), and respective band assignments are given in Table 5.10.

It can be seen in Figure 5.13 (a-c) that majority of loading vectors had featured as heightened biochemical alterations in Figure 5.9(c) and (d). Therefore, the samples discrimination (in Figure 5.12 (a), (c), (e)) were attributed to the heightened biochemical differences between the malignant and the normal samples of the studied patients. The loading plots of PC 3 (4.51%), PC 2 (9.84%) and PC 3 (3.10%) revealed that the changes in tyrosine proteins (850, 854, 860 cm^{-1}), phospholipids (1075, 1078, 1086 cm^{-1}), and nucleic acids (1530, 1532 cm^{-1}) were prominent in blood samples of diseased patients, whereas changes in proteins (745, 752, 755 cm^{-1}), amide III (1240, 1245 cm^{-1}), and tryptophan / phospholipids (1331, 1341, 1343 cm^{-1}) were prominent in blood samples of control patients. Besides, the alterations in phenylalanine (1002 cm^{-1}) and proteins / phospholipids (1126, 1127 cm^{-1}) were prominent in blood samples of control patients in early stages (grade 1, grade 2) of cancer development whereas alterations in saccharides (982, 985 cm^{-1}), tyrosine / phenylalanine proteins (1610, 1615 cm^{-1}), amide I (1643, 1648 cm^{-1}) and nucleic acid base of thymine (1713, 1717 cm^{-1}) were prominent in blood samples of diseased patients in early stages (grade 1, grade 2) of cancer development. Further, changes in CH_2CH_3 bending modes of collagen and phospholipids (1032 cm^{-1}), nucleic acid base of uracil (1502 cm^{-1}), nucleic acid base of cytosine (1695 cm^{-1}) and primary metabolite of citric acid (1733 cm^{-1}) were present during late stages of cancer development (Rehman *et al.*, 2013). This result is in accordance with previous findings (Nargis *et al.*, 2019; Vargas-Obieta *et al.*, 2016), where spectral features at 848 cm^{-1} and 1083 cm^{-1} were found to depict higher Raman intensities in the mean Raman spectra of patient samples, and spectral features around 761 cm^{-1} had higher Raman intensities in spectra of control/healthy volunteers. However, contrary to findings by Vargas-Obieta *et al.*, (2016), the peak at 742 cm^{-1} could be associated with diseased patients rather than healthy volunteers.

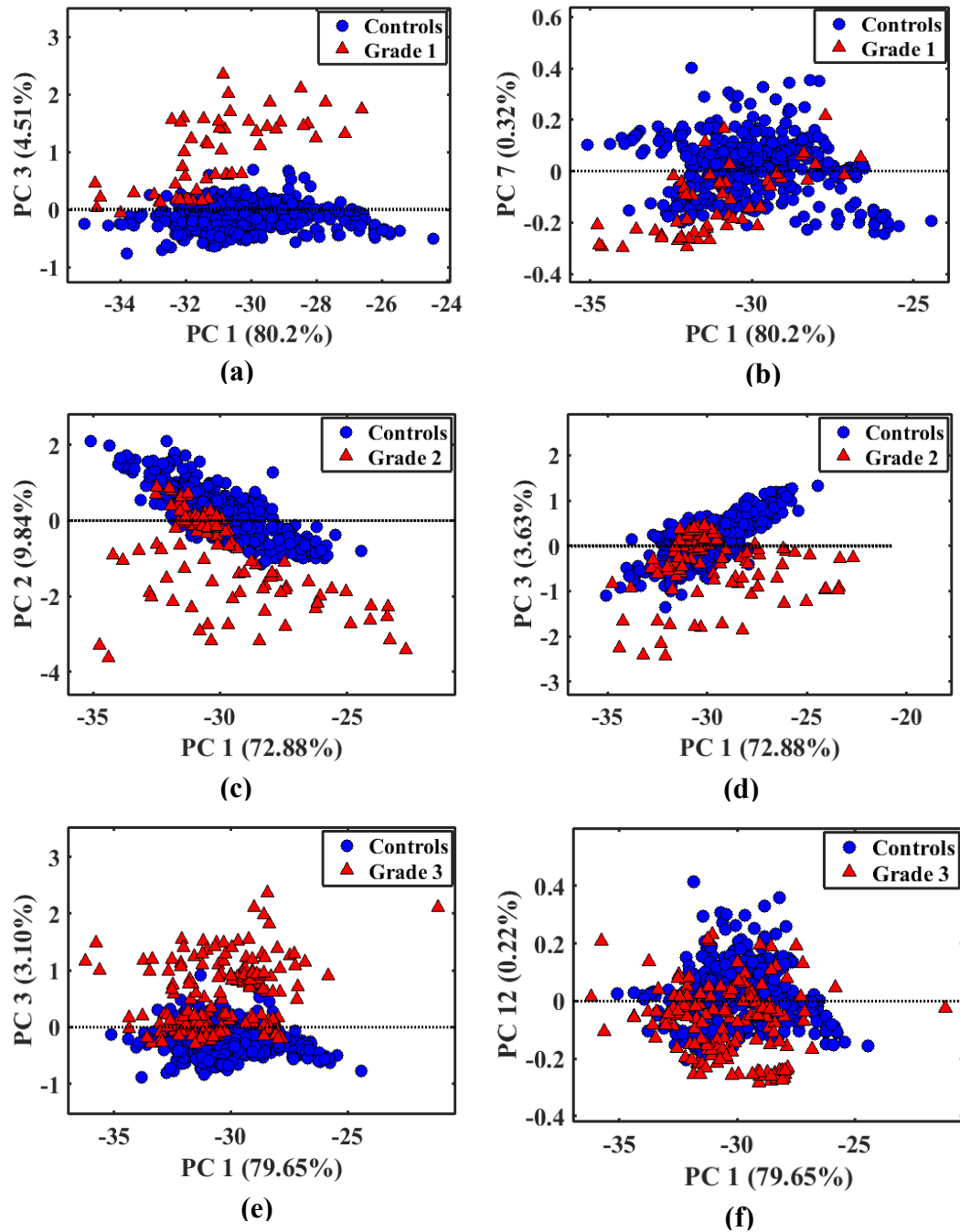


Figure 5.12 Scatter plots showing distribution of the first principal component (PC 1) versus various PCs (2, 3, 7, 12) scores and the diagnostic line for LDA for (a, b) grade₁, (c, d) grade₂, and (e, f) grade₃, spectral datasets, respectively. The overall PCA-LDA classification accuracies for grade 1, 2 and 3 spectral datasets were 100%, 97% and 93%, respectively.

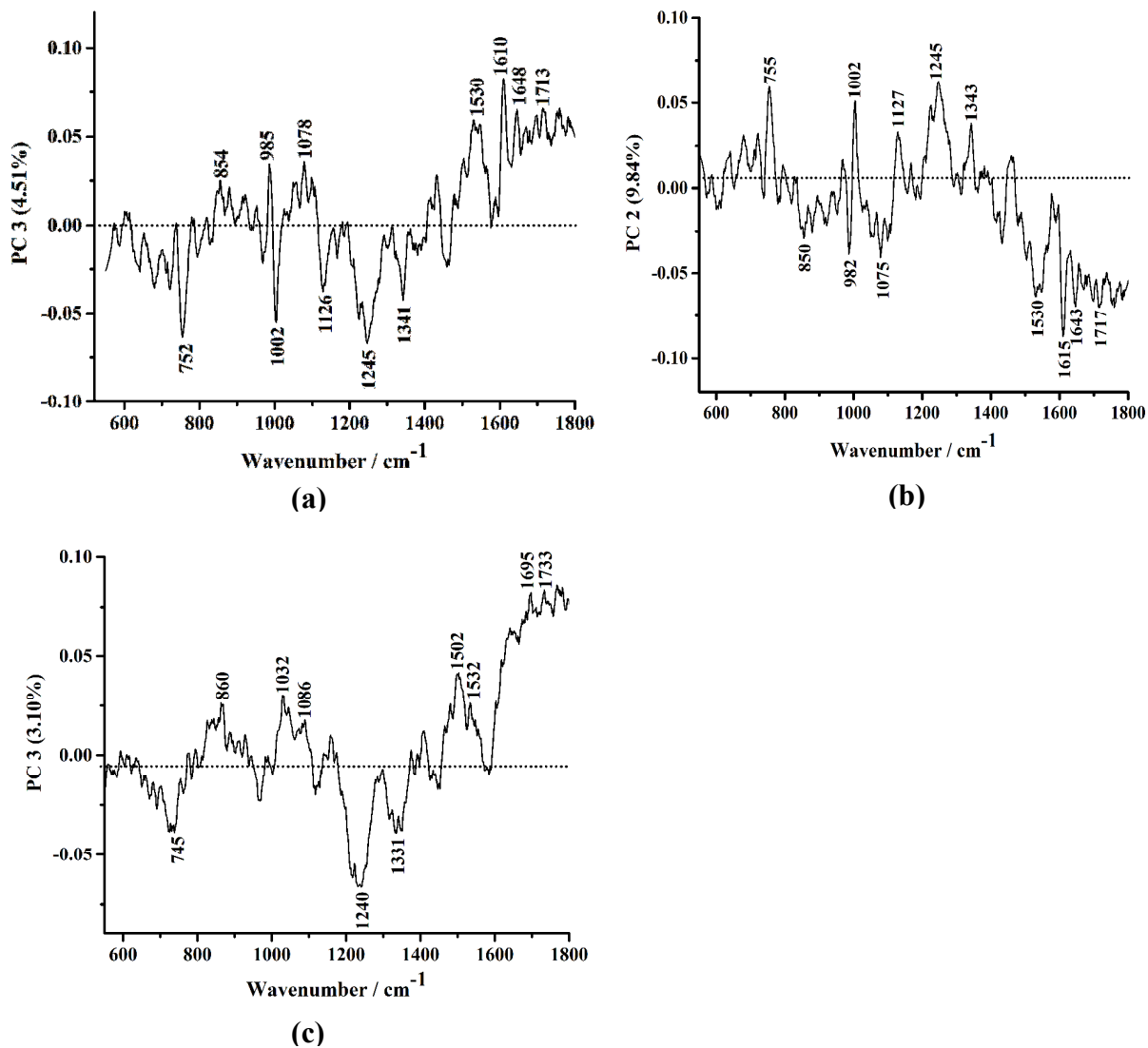


Figure 5.13 Linear discriminant functions showing loading vectors associated with scores discrimination in Figure 5.12 (a, c, e). The loading vectors explain the prominent biochemical alterations differences between healthy and diseased samples.

The LDA discrimination between healthy and diseased scores was further examined using the remaining significant PCs in Table 5.9. Some reasonable level of samples discrimination was observed in loading plots of PCs 7 (0.32%), 3(3.63%), and 12 (0.22%), in grade₁, grade₂, and grade₃ spectral datasets, respectively, as shown in Figure 5.12 b, d, and (f). This was in agreement with their fairly high levels of statistical significance (Table 5.9), and canonical loading parameters (Figure 5.11). The respective loading vectors are shown in Figure 5.14 (a), (b), (c), and band assignments are provided in Table 5.10.

Examination of scatter plots and their respective loading vectors (Figure 5.14) showed the subtle biochemical alterations could be mainly associated with the few extreme vertically end scores on the scatter plane, where the bands around (589, 594, 630, 1160, 1250, 1347, 1358 cm^{-1}) and (858, 868, 1005, 1626, 1630, 1638 cm^{-1}) generally explained the common subtle biochemical alterations in diseased and control (healthy) samples, respectively. By visual inspection, these bands had not been detected in the Raman spectra shown in Figure 5.9 (c) and (d). Therefore, the sample discriminations in Figure 5.12 (b), (d), and (f) can be attributed to the subtle biochemical differences alterations between the malignant and the normal samples of the studied patients. These subtle biochemical differences were mainly due to nucleic acids, proteins, and lipids, where the protein bands were based on aromatic acids of glutamate, phenylalanine, tryptophan, tyrosine, proline, glycine, and valine. In particular, the loading plots of PCs 7 (0.32%), 3(3.63%), and 12 (0.22%) revealed that the subtle changes in glycerol (589 cm^{-1}), tryptophan / phosphatidylinositol (594 cm^{-1}), glutamate / tryptophan (630 cm^{-1}), β -carotene (1160 cm^{-1}), amide III (1250 cm^{-1}), tryptophan / α -helix / phospholipids (1347 cm^{-1}) and tryptophan / guanine (1358 cm^{-1}) were present in blood samples of diseased patients, whereas subtle changes in tyrosine proteins (858 cm^{-1}), proline / saccharides (868 cm^{-1}), phenylalanine (1005 cm^{-1}), glutamate (1626 cm^{-1}), glycine / valine (1630 cm^{-1}) and amide I (α -helix) / β -carotene were present in blood samples of control patients.

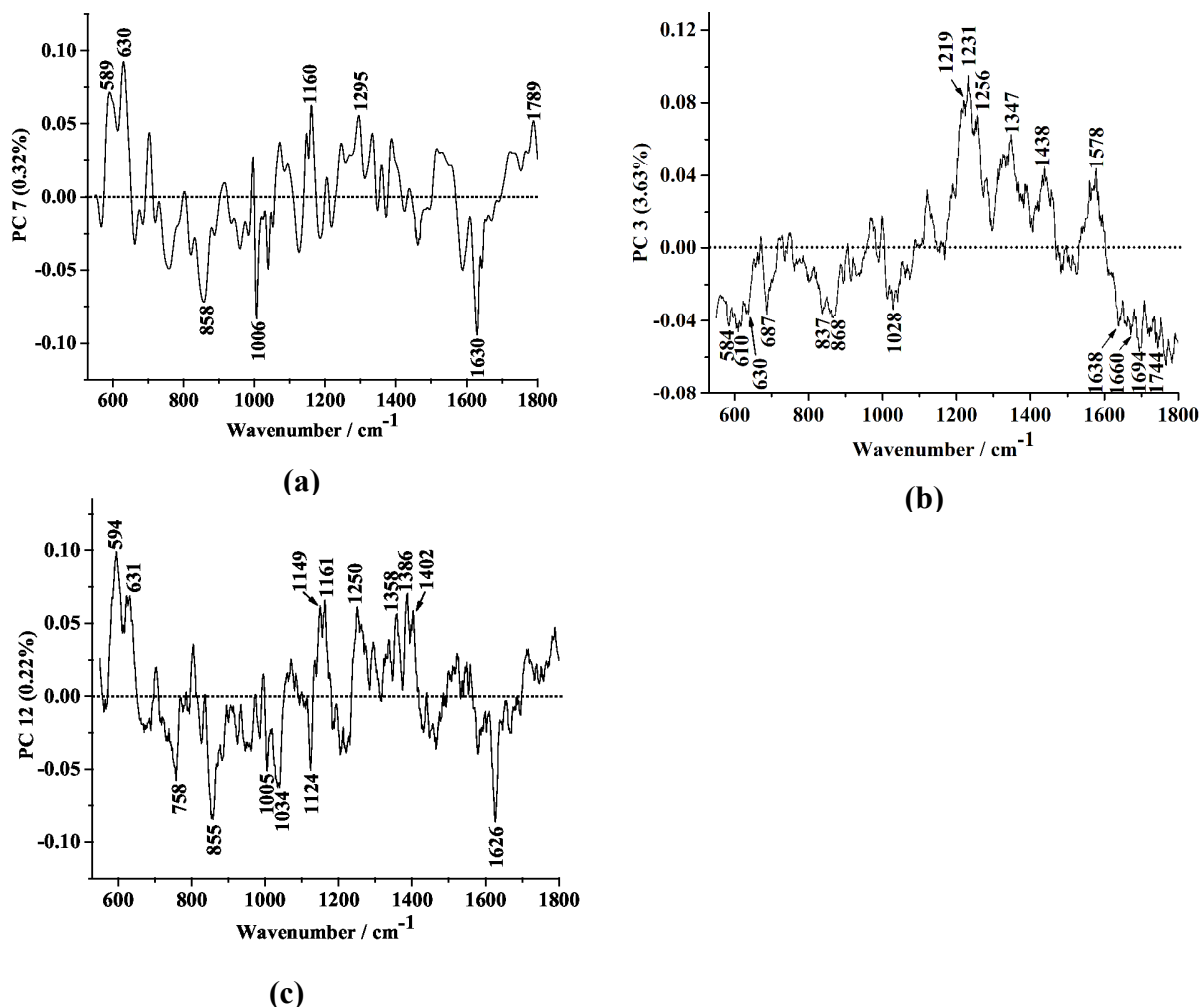


Figure 5.14 Linear discriminant functions showing loading vectors associated with scores discrimination in Figure 5.12 (b, d, f). The loading vectors explain the subtle biochemical alterations differences between healthy and diseased samples. The bands at (858, 868, 1005, 1626, 1630, 1638 cm^{-1}) and (589, 594, 630, 1160, 1250, 1347, 1358 cm^{-1}) explains subtle biochemical alterations in blood samples of controls (healthy) and breast cancer stricken patients, respectively.

Table 5.10 Raman band assignments of alterations in blood samples of healthy and breast cancer patients

| Raman shift (cm^{-1}) | Functional groups and molecular vibration assignments | References |
|-------------------------------------|--|--|
| 589 | Symmetric stretching vibrations, glycerol | (Rehman <i>et al.</i> , 2013) |
| 594 | $\delta(\text{CH}_2 / \text{CH}_3)$ deformations (phosphatidylinositol), and tryptophan proteins, | (Gelder <i>et al.</i> , 2007) |
| 630 | $\nu(\text{C-S})$, glutamate, C-C twisting modes of tryptophan | (Chandra <i>et al.</i> , 2015) |
| 745 | C-S stretch of phospholipids, ring breathing modes of DNA / RNA bases (thymine) | (Pichardo-Molina <i>et al.</i> , 2007) (Vargas-Obieta <i>et al.</i> , 2016) |
| 752, 755 | Symmetric breathing of tryptophan | (Pichardo-Molina <i>et al.</i> , 2007) |
| 850 | Ring breathing modes of tyrosine proteins | (Vargas-Obieta <i>et al.</i> , 2016) |
| 854, 860 | Ring breathing modes of tyrosine proteins | (Pichardo-Molina <i>et al.</i> , 2007) |
| 868 | C-C stretch of proline, C-O-C skeletal mode of saccharides) | (Chandra <i>et al.</i> , 2015) |
| 982, 985 | C-C stretching, β -sheet (proteins), C-O-C skeletal mode of saccharides | (Rehman <i>et al.</i> , 2013) |
| 1002, 1005 | Symmetric ring breathing mode of phenylalanine | (Pichardo-Molina <i>et al.</i> , 2007) |
| 1075, 1078 | C-C and C-O stretch of phospholipids | (Rehman <i>et al.</i> , 2013) |
| 1086 | O-P-O and C-C stretch of phospholipids | (Vargas-Obieta <i>et al.</i> , 2016) |
| 1126, 1127 | C-N stretch (proteins), ν (C-C) stretch of phospholipids | (Pichardo-Molina <i>et al.</i> , 2007) |
| 1160 | C-C / C-N stretching (proteins), ring breathing modes of DNA / RNA bases (adenine, guanine) | (Gelder <i>et al.</i> , 2007) |
| 1240, 1245 | Asymmetric phosphate vibrations, β – sheet (amide III), CH_2 (glycine and proline) | (Rehman <i>et al.</i> , 2013) |
| 1250 | β – sheet (amide III collagen assignment) | (Vargas-Obieta <i>et al.</i> , 2016) |
| 1331 | C-N stretch of tryptophan proteins, α -helix, and phospholipids | (Movasaghi <i>et al.</i> , 2007) |
| 1341, 1343 | C-N stretch of tryptophan, α -helix, phospholipids | (Vargas-Obieta <i>et al.</i> , 2016) |
| 1347 | C-N stretch of tryptophan, α -helix, phospholipids | (Movasaghi <i>et al.</i> , 2007) |

Table 5.10 (continued) Raman band assignments of alterations in blood samples of healthy and breast cancer patients

| Raman shift (cm^{-1}) | Functional groups and molecular vibration assignments | References |
|-------------------------------------|--|--|
| 1358 | C-N stretch of tryptophan, ring breathing modes of DNA / RNA bases (guanine) | (Chandra <i>et al.</i> , 2015) |
| 1530, 1532 | Ring breathing modes of DNA / RNA bases (adenine, guanine, cytosine) | (Rehman <i>et al.</i> , 2013) |
| 1610, 1615 | NH ₂ bending of tyrosine, phenylalanine | (Pichardo-Molina <i>et al.</i> , 2007) |
| 1626 | C=C stretch, glutamate | (Gelder <i>et al.</i> , 2007) |
| 1630 | C=C stretch, glycine, valine | (Rehman <i>et al.</i> , 2013) |
| 1638 | Amide I (α -helix), β -carotene | (Rehman <i>et al.</i> , 2013) |
| 1643, 1648 | Amide I (α -helix) | (Chandra <i>et al.</i> , 2015) |
| 1695 | Amide I (α -helix) | (Gelder <i>et al.</i> , 2007). |
| 1713, 1717 | C=O stretch of DNA / RNA bases (thymine) | (Movasaghi <i>et al.</i> , 2007) |

5.2.1.3 Quantitative analysis of trace biomarkers in whole blood spectra using partial least-squares regression

First, we examined the Raman spectra of the pure biochemical components that were chosen to represent the major biochemical groups in cellular constituents which include nucleic acids, proteins, lipids, amino acids and polysaccharides. Figure 5.15 shows the Raman spectra of selected basic pure biochemical components, i.e., the bovine serum albumin, glycogen type IX from bovine liver, glycerol trioleate derived from glycerol, triolein, L-glutamic acid potassium salt monohydrate, glycine and RNA extract that were used for preparation of calibration samples (spectra have been linearly offset for clarity). The observed bands for albumen, glycogen, glutamate, glycine, RNA and triolein biochemical components included (856, 1001, 1244, 1444, 1653 cm^{-1}), (853, 867, 937, 1053, 1120, 1334 cm^{-1}), (664 cm^{-1}), (602, 890, 1321 cm^{-1}), (653, 929, 1009, 1061, 1344 cm^{-1}), and (831, 1141, 1227, 1658 cm^{-1}) respectively, which agreed with expected biochemical assignments in literature (Gelder *et al.*, 2007; Rehman *et al.*, 2013). The high intensity peak at 1009 cm^{-1} in RNA spectrum is assigned to cytosine (Gelder *et al.*, 2007). Although RNA extract was assumed to have DNA contamination, it was noted the common band

associated with ring breathing modes of adenine e.g., the 785 cm^{-1} marker, commonly seen in in DNA spectra could not be detected.

The model fitting was evaluated by correlating the reference values and the values calculated by the models of the prediction set. It was observed that 8 optimal components accounting for total cumulative variance of 98.99% yielded the best model. The predicted versus measured regression plots showing how the PLS model predicted concentration levels for the calibration samples and how well the model could be expected to perform during the quantification of new similar matrix composition are shown in Figure 5.16. It can be observed that the model worked well for RNA, triolein, glycerol, glycogen, and albumen components. However, the model did not work well for the glycine component ($R_{val}^2 < 0.8$), and the glycine component was therefore discarded during the quantification study. The limits of detection for biochemical compounds in simulate whole blood samples are summarized in Table 5.11. It is observed the biochemical components were quantifiable using the trained PLS regression model and the detection limits lie in the range of calibration set. LOD values suggested there were adequate analyte concentration present to yield an analytical signal that could be well measured from analytical noise, whereas LOQ demonstrated quantitative results could be obtained with a specified degree of confidence (Taleuzzaman, 2018). Besides, the low RMSEP values calculated by summing all squared prediction errors during cross-validation suggested higher reliability and predictive ability of the model (Gontijo *et al.*, 2014), which was verified by the corresponding higher R^2 values. The accuracy and reliability of the PLS regression model was assessed by analyzing concentration levels of a standard laboratory reference material (simulate blood fluid) spiked with known concentrations of biochemical components (Table 5.12). It can be seen that the biochemical components levels were in agreement with known values in a typical standard sample in the range of $\leq 10\%$. The relative amounts of biochemical components (ppm) were calculated by fitting the basal spectra in spectral datasets of the 13 spectra markers measured from blood samples at $589, 594, 630, 858, 868, 1005, 1160, 1250, 1347, 1358, 1626, 1630,$ and 1638 cm^{-1} , and their results are summarized in Table 5.13.

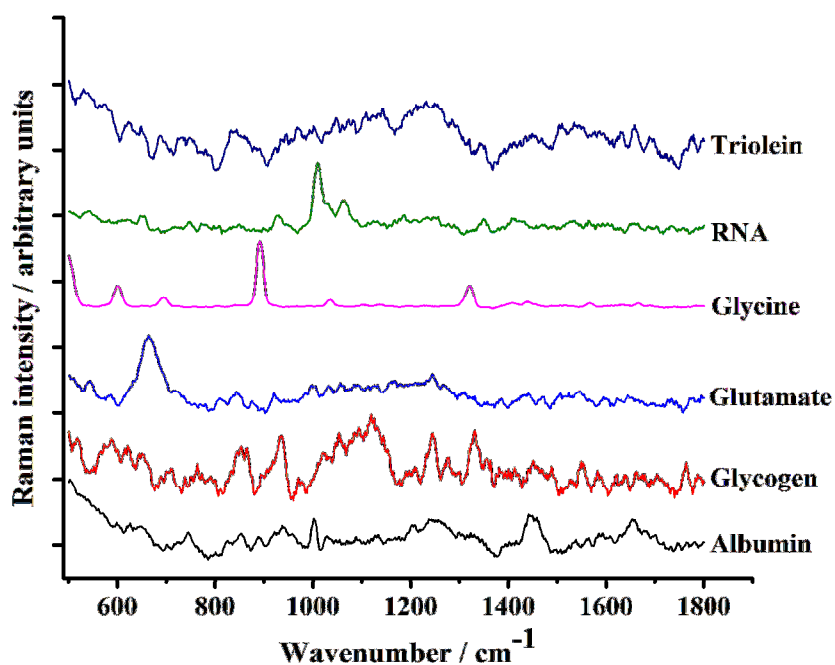


Figure 5.15 Mean Raman spectra of the biochemical constituents used in the concentration fit. The spectra have been linearly offset for clarity.

Table 5.11 Detection limits of biochemical components for Raman analysis of simulate blood fluid

| Biochemical component | Detection limits (mg/ml) | | | |
|-----------------------|--------------------------|---------------------|-----------------------|--------|
| | LOD | LOQ | (RMSEP) | R^2 |
| Albumen | 0.0144 | 0.0438 | 0.00168 | 0.9934 |
| Glycogen | 0.018 | 0.056 | 0.00201 | 0.988 |
| Glutamate | $1.683 \cdot 10^{-8}$ | $5.1 \cdot 10^{-8}$ | $1.825 \cdot 10^{-9}$ | 1 |
| Glycerol | 0.0055 | 0.0169 | 0.000605 | 0.998 |
| RNA | 0.00136 | 0.0041 | 0.00014 | 0.999 |
| Triolein | 0.0506 | 0.153 | 0.0056 | 0.908 |

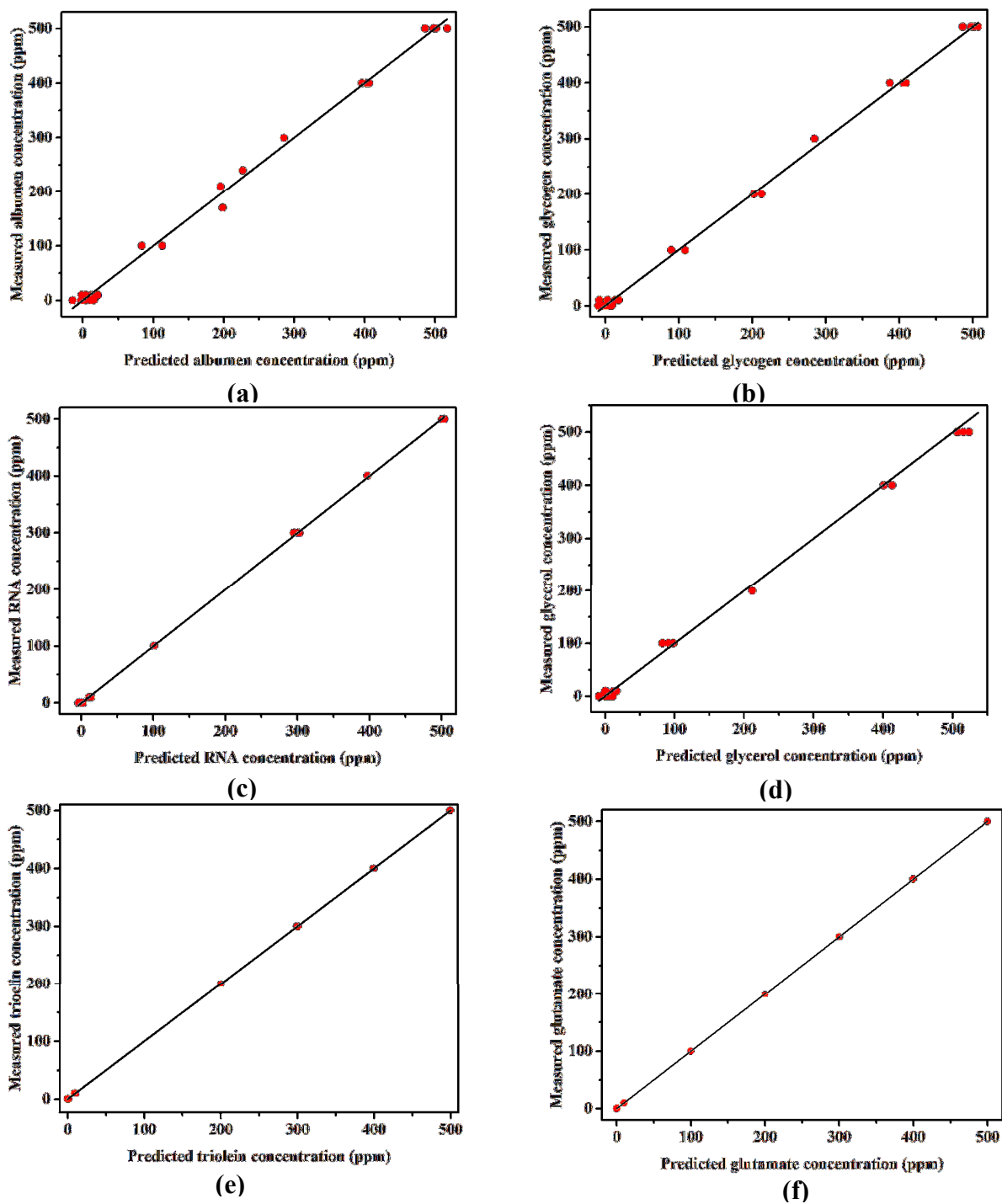


Figure 5.16 Regression plots for partial least squares measured versus predicted biochemical concentrations of the basal compounds (albumen, glycogen, RNA / DNA, glycerol, triolein and glutamate) used in the spectral model, based on the spectra profiles of 13 spectra markers, i.e., 589, 594, 630, 858, 868, 1005, 1160, 1250, 1347, 1358, 1626, 1630, and 1638 cm^{-1} .

Table 5.12 Comparison of biochemical components concentrations in a whole blood simulate reference solution and the results obtained from PLS regression

| Biochemical Components | Concentration (mg / ml) | Measured value (\pm SD) | Deviation (%) |
|------------------------|-------------------------|----------------------------|---------------|
| Albumen | 0.4 | 0.37 \pm 0.02 | 7.5 |
| Glycogen | 0.1 | 0.094 \pm 0.025 | 6 |
| Glutamate | 0.001 | 0.000953 \pm 0.00012 | 4.7 |
| Glycerol | 0.01 | 0.0108 \pm 0.0024 | 8 |
| RNA | 0.002 | 0.0021 \pm 0.00011 | 5 |
| Triolein | 0.3 | 0.273 \pm 0.0036 | 9 |

Table 5.13 Relative amounts of biochemical components in blood samples of healthy and breast cancer patients- based on the determined trace biomarker alterations

| Disease status | Biochemical components (ppm) | | | | | |
|----------------|------------------------------|----------|-----------|----------|------|----------|
| | Albumen | glycogen | glutamate | glycerol | RNA | triolein |
| Controls | 233.86 | 73.7 | 10.48 | 190 | 62.1 | 18.0 |
| Stage 1 | 237.82 | 98.0 | 60.49 | 234 | 66.4 | 71.95 |
| Stage 2 | 286.03 | 36.4 | 83.69 | 271 | 68.9 | 99.73 |
| Stage 3 | 384.96 | 84.3 | 14.31 | 297 | 96.8 | 101.2 |

For plotting, the determined concentration levels of trace biomarker alterations were normalized to their mean intensities and their levels to cancer presence and severity compared (Figure 5.17). It can be seen that the relative amounts of albumen, triolein, glycerol, and RNA increased with disease status ($p < 0.05$). To the best of our knowledge, there has not been previous spectroscopic studies comparing biochemical alterations levels in blood samples of healthy (controls) and breast cancer patients, based on few selected spectral regions (589, 594, 630, 858, 868, 1005, 1160, 1250, 1347, 1358, 1626, 1630, and 1638 cm^{-1}) as demonstrated in this study. Nevertheless, although this study was limited in the number of samples in each diseased stage, the observed differences in relative amounts of biochemical components gives insight into possible pathological differences during breast cancer progression.

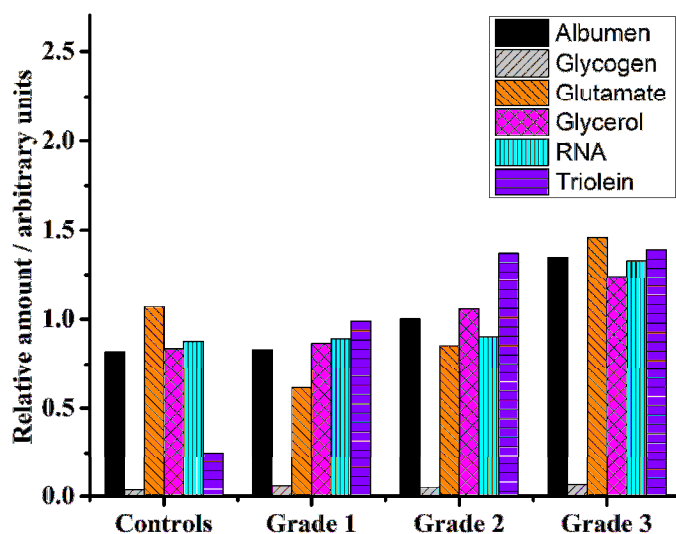


Figure 5.17 Plot of the relative contribution of selected basal compounds estimated by the Raman spectral model applied to the spectra of blood samples from the controls, grade 1, grade 2, and grade 3 breast cancer patients. For plotting, the relative amounts of components are normalized to their mean amount value.

By using $600\text{-}1800\text{ cm}^{-1}$ spectral region for characterizing different stages of breast cancer, Nagris *et al.*, (2019) found that nucleic acids and proteins were exclusively present in patient plasma patients. Similarly, we also performed stage wise comparison for breast cancer in $600\text{-}1800\text{ cm}^{-1}$ spectral region using the band ratios (i.e., I_C/I_N) (not shown) where we found changes in nucleotides (689 cm^{-1}), anti-symmetric phosphate vibrations (1185 cm^{-1}), phospholipids (1285 cm^{-1}) and guanine (1319 cm^{-1}) increased with malignancy. In the current context, the observed increase in RNA content with disease status suggests heightened protein synthesizing activity in cells, and therefore potential increase in nuclear contents.

The increase of albumen component suggested blood samples from diseased patients had higher levels of proteins content. In this study, utility of selected spectral regions showed the changes in tryptophan ($594, 628\text{-}632\text{ cm}^{-1}$), glutamate ($628\text{-}632, 1626\text{ cm}^{-1}$), tyrosine ($858\text{-}860\text{ cm}^{-1}$), glycine / valine (1630 cm^{-1}) and amide I (α -helix (1638 cm^{-1}) increased with malignancy. It should be noted that intense spectral regions of proteins have been previously observed in blood samples (Nargis *et al.*, 2019) and breast tissues (Chowdary *et al.*, 2006) of breast cancer patients, when compared to those from control patients. In Chowdary *et al.*, (2006), spectral differences analysis between normal, benign and malignant breast tissues in $800\text{ - }1800\text{ cm}^{-1}$ region showed normal and (benign, malignant) tissues were dominated by lipids ($1078, 1267, 1301, 1440, 1654,$

1746 cm^{-1}) and proteins (stronger amide I, red shifted ΔCH_2 , broad and strong amide III, 1002, 1033, 1530, 1556 cm^{-1}) alterations respectively. This dominance of protein biochemical alterations was also reported by Gonzálezsolís *et al.*, (2011), where tryptophan protein and amide III (at 1244 cm^{-1}) biomolecules alterations were observed in late stages of breast cancer progression. The increase in protein concentration can be generally attributed to oxidative stress associated with breast cancer progression (Marinello *et al.*, 2014).

With reference to increase in triolein and glycerol components, our analysis based on selected spectral regions suggested the whole blood samples of diseased patients had relatively more cell fat than in healthy patients, which was an indication of higher metabolic activity in malignant samples. In particular, changes in glycerol (589 cm^{-1}) and phosphatidylinositol (594 cm^{-1}) were found to increase with malignancy. From histochemistry perspective, this may be explained by the fact that lipid-mobilizing effect of the tumour may be necessary in sustaining tumour growth (Mulligan *et al.*, 1991). Literature concerning spectroscopic comparison of lipid alterations levels in blood samples of healthy (control) and diseased breast cancer patients is scarce, and even the limited tissue-based studies have previously reported mixed results. For instance, a study by Chowdary *et al.*, (2006) showed normal breast tissues had higher levels of lipids when compared to benign and malignant breast tissues, while malignant tissues contained relatively more lipids in comparison to benign tissues. Elsewhere, a spectral analysis study by Rehman *et al.*, (2010) based on 1522, 1540, 1630 and 1640 cm^{-1} spectral regions showed that dominance of acylglyceride, and proteins in higher and low- nuclear-grade spectrum, respectively. Blood serum and blood plasma-based Raman spectroscopy studies aimed at understanding lipid-mobilizing effects of breast tumors are highly encouraged.

PLS-DA scatterplots showing differentiation of stages of cancer are detailed in Figure 5.18. The respective significant latent variables (loadings) are shown in Figures 5.19 and 5.20. In classifying healthy (controls) scores against grade 1 breast cancer scores (see Figure 5.19 (a)), the loading plots shows negative loadings had higher intensities at 593 cm^{-1} (tryptophan / phosphatidylinositol) and 631 cm^{-1} (glutamate / tryptophan), meaning control samples had higher alterations at these bands, whereas positive loadings had higher intensities at 1357 cm^{-1} (tryptophan / guanine), 1626 cm^{-1} (glutamate) and 1630 cm^{-1} (glycine, valine) which can be associated with breast cancer stage-1 malignancy. If we consider (see Figure 5.19 (b)), the differentiation of healthy (controls) scores against grade 2 breast cancer scores showed control scores had higher loadings at 594 cm^{-1} (tryptophan / phosphatidylinositol) and 632 cm^{-1} (glutamate

/ tryptophan) bands whereas grade 2 breast cancer scores had intense loadings at 1003 cm^{-1} (phenylalanine) 1624 cm^{-1} (glutamate) and 1636 cm^{-1} (amide I (α -helix)/ β -carotene). In Figure 5.19 (c), positive loadings had higher intensities in the Raman spectral data of healthy (control) patients than stage-3 patients, which included those at 1248 cm^{-1} (amide III) and 1349 cm^{-1} (tryptophan / α -helix / phospholipids), whereas negative loadings had higher intensities in the Raman spectral data of grade 3 breast cancer patients than healthy (control) patients, which included those at 594 cm^{-1} ((tryptophan / phosphatidylinositol) 630 cm^{-1} (glutamate / tryptophan), 1626 cm^{-1} (glutamate), 1630 cm^{-1} (glycine / valine) and 1638 cm^{-1} (amide I (α -helix)/ β -carotene).

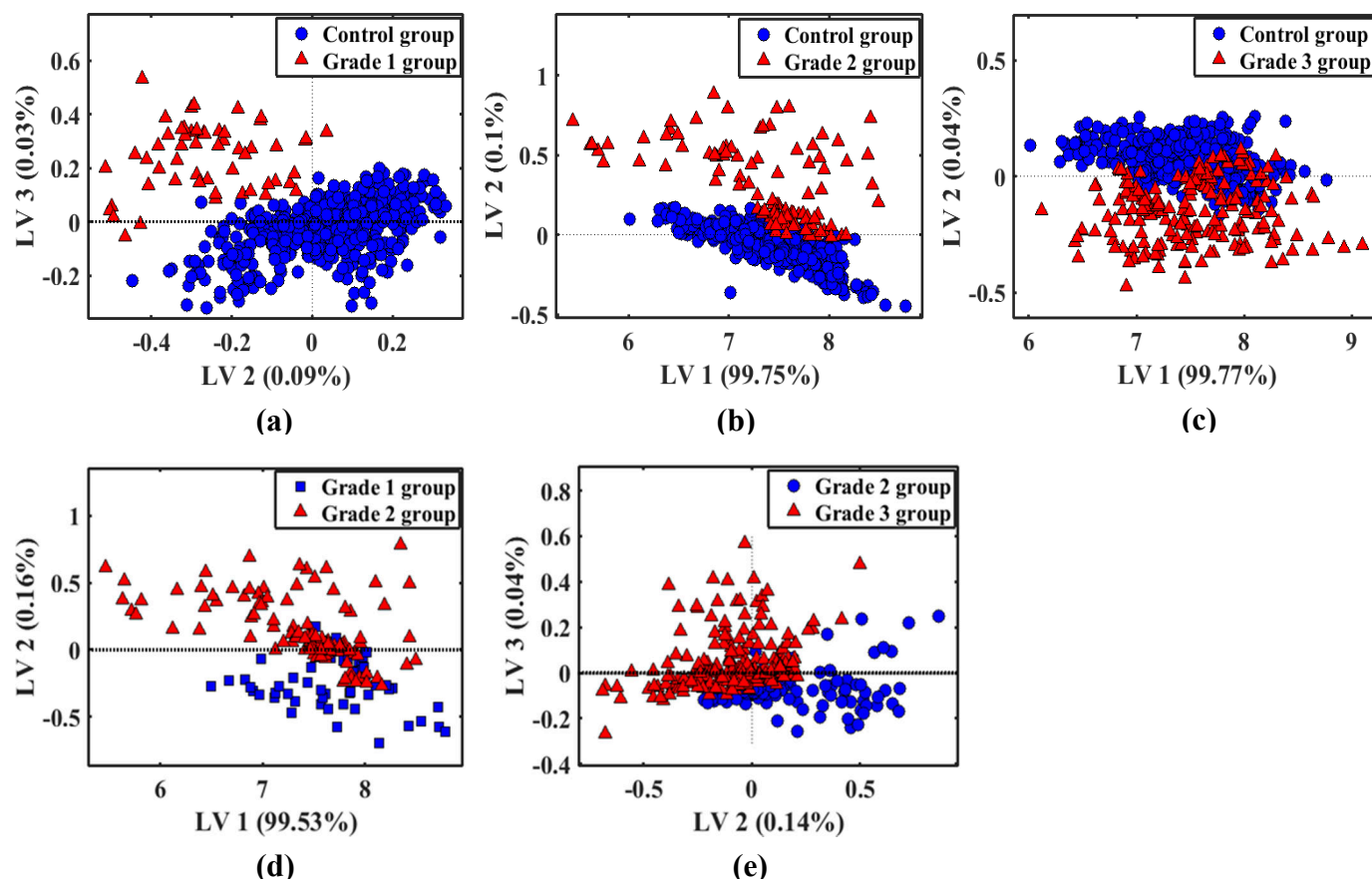
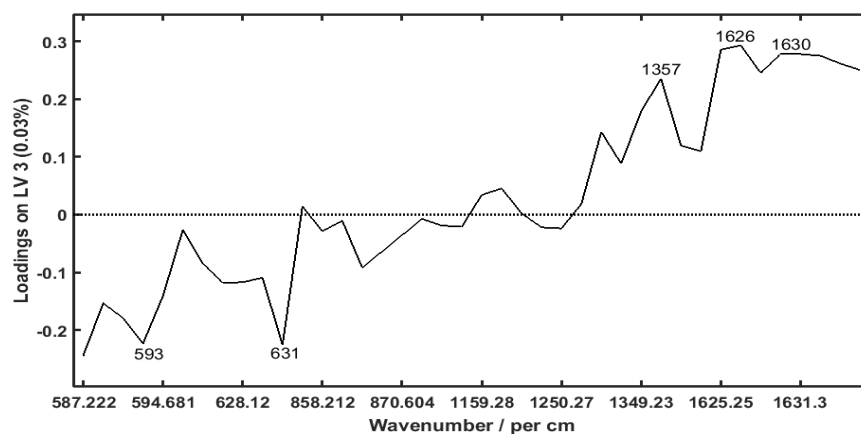


Figure 5.18 PLS-DA scatterplots showing differentiation of spectra of healthy samples, grade 1, grade 2, and grade 3 breast cancer samples.

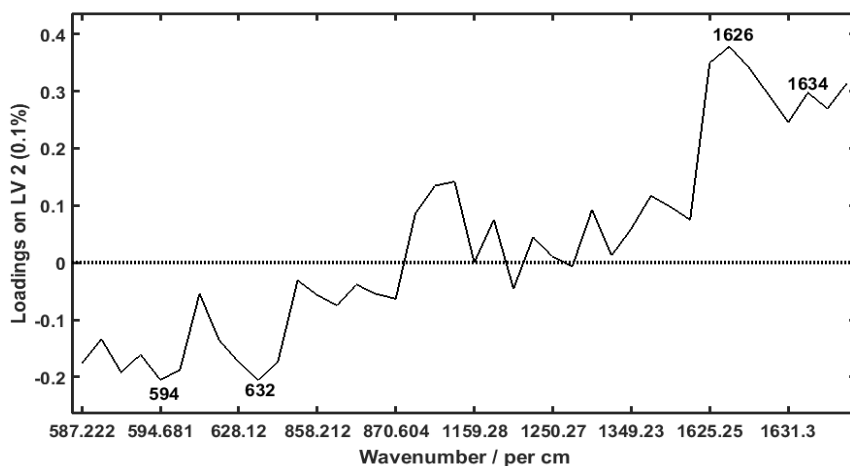
After the diagnosis of malignancy, it is important to determine the staging of the cancer. This will help to know that how far the disease has progressed and also to determine the best way to contain and eliminate the breast cancer. In this regard, Raman spectroscopy can potentially be of benefit in the differentiation of these stages of breast cancer, hence leading to early diagnosis which could be useful for the effective treatment. The progress of cancer was studied by staging levels of cancer i.e. breast cancer grade-1 versus breast cancer grade-2, and breast cancer grade -2 vs breast cancer grade -3. As seen in Figure 5.20 (a), the negative loadings represent the Raman spectral features (596, 630, 860 cm^{-1}) that had higher intensities in breast cancer grade -1 patients whereas positive loadings represent the Raman spectral features (1003, 1245, 1626, 1637 cm^{-1}) that have higher intensities breast cancer grade -2 patients. Comparison of breast cancer grade -2 and breast cancer grade -3 scores was performed by examining latent variable / loading 2 (LV 2) and latent variable / loading 3 (LV 3), as shown in Figure 5.20 (b)-(c). It can be seen that the Raman spectral features associated with breast cancer grade -2 included 590, 596, 620, 630 / 632, 860, 1001 and 1348 cm^{-1} , while spectral features at 1006, 1249, 1347, 1358, 1626 / 7, 1630 and 1637 cm^{-1} indicated the heightened biochemical changes during late breast malignancy.

To better understand these differences, band ratios i.e., I_C/I_N were calculated by dividing the normalized intensities of diseased spectra (I_C) by normalized intensities of control spectra (I_N) at the determined subtle bands. The ratio values at 589, 594 (tryptophan, phosphatidylinositol), 628-632 (tryptophan, glutamate), 858-860 (tyrosine), 1158-1163 (nucleic acids) and 1626-1638 (amide I) were found to increase with malignancy. The band ratios of intensities at the other subtle bands demonstrated inconsistent trend across all levels of stage development.

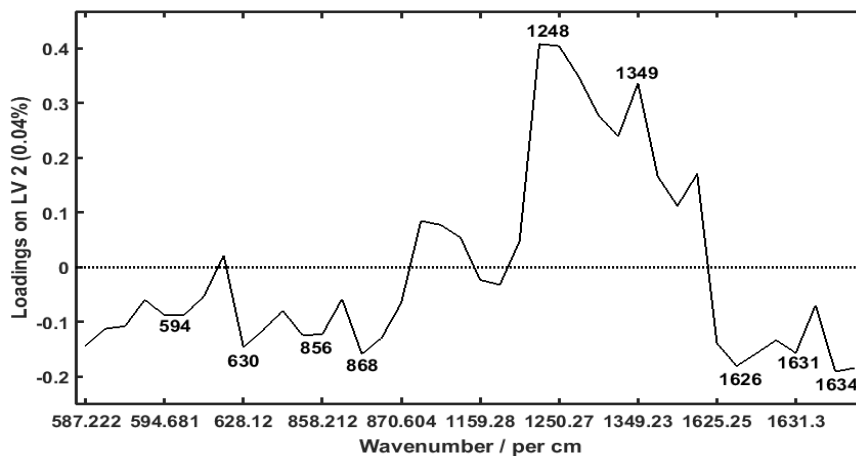
The accuracy, sensitivity and specificity classification parameters of PLS-DA diagnostic model were analyzed (Table 5.14). The number of correctly identified cases out of total cases led to an accuracy of 98%, 98% and 94% for grade -1, grade -2 and grade -3 cancers, respectively. The sensitivity, expressed as the number of correctly identified cancer spectra over the total number of diseased spectra was found to be 100% for grade 1 cancer, 98% for grade 2 cancer and 94% for grade 3 cancer. The specificity, expressed as the number of correctly identified healthy (control) spectra over the total number of healthy spectra was determined to be >96%, for all stages of cancer.



(a)

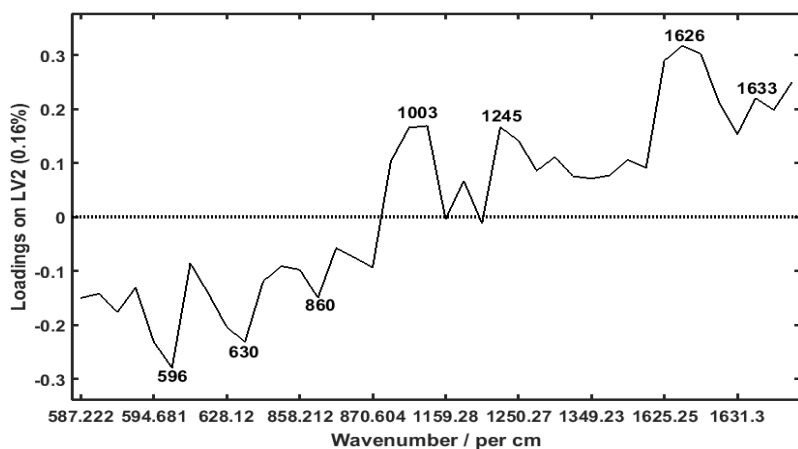


(b)

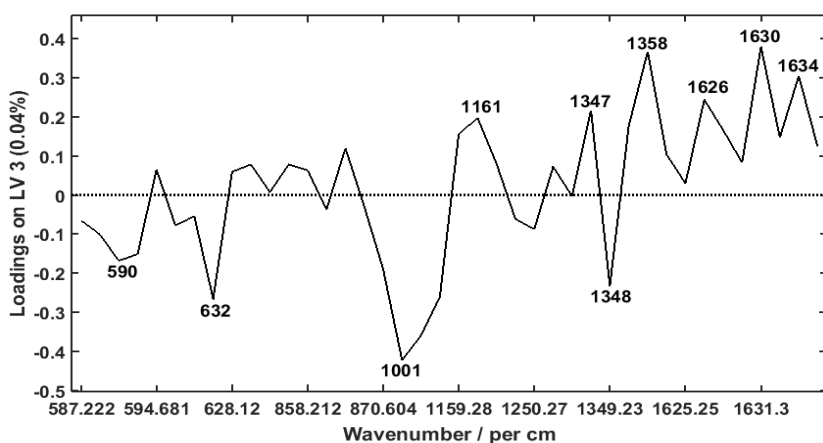


(c)

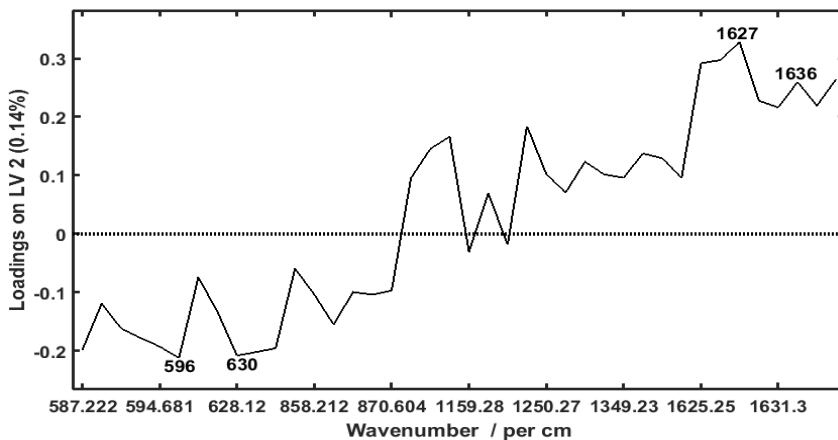
Figure 5.19 The loading vector plots that explain differentiation of (a) controls versus grade 1 breast cancer, (b) controls versus grade 2 breast cancer, and (c) controls versus grade 3 breast cancer.



(a)



(b)



(c)

Figure 5.20 The loading vector plots that explain differentiation of (a) grade 1 versus grade 2 breast cancer, and (b), (c) grade 2 versus grade 3 breast cancer.

Table 5.14 Diagnostic results of PLS-DA on the Raman spectra of whole blood from healthy volunteers (controls) and breast cancer patients based on the selected spectral regions (589, 594, 630, 858, 868, 1005, 1160, 1250, 1347, 1358, 1626, 1630, and 1638 cm^{-1}).

| Cases | | | | | | | |
|----------------|---------------|---------------|----------|-------|----------|-------------|-------------|
| Disease status | Diagnosis | Breast cancer | Controls | Total | Accuracy | Sensitivity | Specificity |
| Grade-1 | Breast cancer | 50 | 5 | 55 | 98% | 100% | 99% |
| | Controls | 6 | 433 | 439 | | | |
| Grade-2 | Breast cancer | 107 | 0 | 107 | 98% | 98% | 100% |
| | Controls | 8 | 427 | 435 | | | |
| Grade-3 | Breast cancer | 169 | 21 | 190 | 94% | 89% | 96% |
| | Controls | 17 | 418 | 435 | | | |

The results obtained in this study strengthen the view that higher principal components can be useful in extracting weak band alterations. For instance spectral markers at (853, 854, 858 cm^{-1}), (1002, 1003 cm^{-1}) and 1156 cm^{-1} had been previously identified in works by Vargas-Obieta *et al.*, (2016), Bilal *et al.*, (2017), Vargas-Obieta *et al.*, (2016) and Bilal *et al.*, (2017), respectively. Although our work could not optimally detect same spectra markers (see Figure 5.9(b-d)), the utility of higher principal components in extracting weak variance signals yielded spectral markers at 858 cm^{-1} , 1003 cm^{-1} and 1160 cm^{-1} , a disparity which can be attributed to different experimental conditions. Our study also advances findings of previously related studies (Nargis *et al.*, 2019; Khanmohammadi *et al.*, 2010; Pichardo-Molina *et al.*, 2007; Vargas-Obieta *et al.*, 2016; Bilal *et al.*, 2017) in that it identifies biochemical alterations at 589 (glycerol), 594 cm^{-1} (tryptophan, phosphatidylinositol) and 630 cm^{-1} (glutamate, tryptophan), 1626 (glutamate), 1630 (glycine / valine), and 1638 (amide I (α -helix), β -carotene), which have not been previously reported.

The PLS-DA was found to perform well in diagnosing and staging of breast malignancy when compared to PCA-LDA. This difference is due to the manner in which both algorithms handle the datasets. Different from PCA that consider spectra matrix as one set of data, the PLS realizes dimensionality reduction by considering the relations between two data blocks (e.g., X and Y) across the same samples (Liu *et al.*, 2016). Consequently, PLS maximizes the covariance between X and Y, thus explaining much variance in the datasets (Liu *et al.*, 2016). In our study, diagnosis of breast malignancy using PLS-DA model yielded classification accuracies, sensitivity

and specificity values of >90%, indicating that combination of PLS-DA and Raman spectroscopy is a potential tool for cancer diagnostics in body fluids.

5.2.1.4 Multivariate exploratory analysis of Independent Component Analysis (ICA), Multidimensional Scaling (MDS), and Partial least Square Discriminant analysis (PLS-DA) for breast cancer diagnostics in blood

Identification of every biochemical component in a complex mixture such as that found in a cell or tissue may not be possible, because components present exist in many different forms each having a slight different Raman spectrum (Shafer-peltier *et al.*, 2002). To obtain a more comprehensive picture of the chemical component in their microenvironment within normal or diseased tissue, ICA can be used as a pattern recognition algorithm to extract information from Raman spectra due to its ability of providing information with statistical independency (Bouzalmat & Kharroubi, 2014). ICA by Maximum Likelihood (ML) fast fixed-point estimation algorithm was applied on matrix created from the dataset of the 13 spectral regions: 589, 594, 630, 858, 868, 1005, 1160, 1250, 1347, 1358, 1626, 1630, and 1638 cm^{-1} . These spectral regions were attributed to proteins, lipids, and nucleic acids components, where the protein bands were based on aromatic acids of glutamate, phenylalanine, tryptophan, tyrosine, proline, glycine, and valine (Table 5.10). By looking for maximum likelihood, independent information were separated from the spectra. Among separated information, the high frequency noise containing information of baseline shift; perhaps generated by the ambient noise and low frequency noise were observed. Before applying ICA, all spectra were mean centered in order to enhance the peak information, whitened and preprocessed to have unit variance in order to shorten calculation time (Yao *et al.*, 2012). The processing time was very short (< 10 seconds) because fast fixed point algorithm was added to ML estimation method.

Figure 5.21 shows the eigenvalues that were determined by maximum likelihood estimation on blood samples. In general, the first eigenvalue was found to be the most intense and accounted for much variance during whitening process. For all datasets, 10 eigenvalues that accounted for more than 90% variance were selected for further analysis (Table 5.15). It can be seen (Table 5.15) that the sum of eigenvalues (in percentage) for the number of retained eigenvalues decreased with stage of cancer progression, meaning there were some additional spectral regions in respective datasets that were characteristic of noise, that could not be useful for cancer diagnosis (Crow *et al.*, 2005). However, to determine if corresponding coefficients of the

combinations decomposed by ICA were useful for diagnosis, the PLS-DA algorithm was included for classification. For PLS-DA, 3 latent variables and 10-fold cross-validation groups were found appropriate for discriminating control from diseased scores. The respective ICA-PLSDA scatter plots for (a) grade 1, (c) grade 2, and (e) grade 3 datasets of whole blood samples from normal and breast cancer are shown in Figure 5.22.

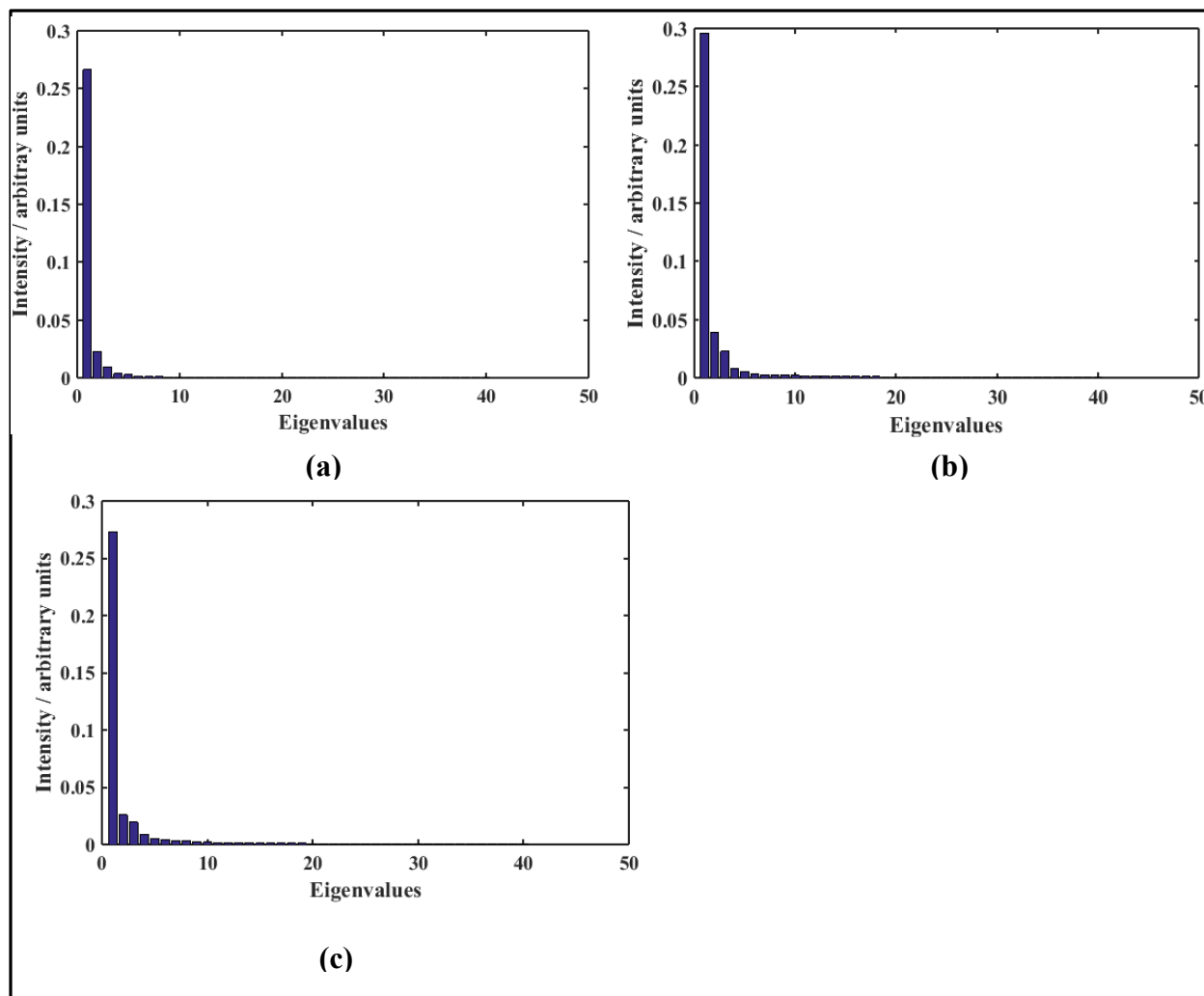


Figure 5.21 The ICA eigenvalues for Raman spectra of blood samples from healthy volunteers and (a) grade 1, (b) grade 2, and (c) grade 3 breast cancer patients. The analysis is performed for spectral datasets at 13 spectral bands: 589, 594, 630, 858, 868, 1005, 1160, 1250, 1347, 1358, 1626, 1630, and 1638 cm^{-1} .

Table 5.15 Selected dimensions (eigenvalues) and respective explained total variances for ICA by Maximum Likelihood (ML) fast fixed-point estimation on Raman spectra of blood samples from healthy volunteers (control) and breast cancer patients

| Datasets | Dimensions | Sum of eigenvalues retained (%) |
|----------|------------|---------------------------------|
| Grade 1 | 10 | 96.61% |
| Grade 2 | 10 | 94.53% |
| Grade 3 | 10 | 93.99% |

As observed in Figure 5.22, majority of diseased scores were clustered in a centroid region, meaning they shared common biochemical characteristics and spectral regions associated with independent components (ICs) were statistically independent. For grade 1 dataset, the fifth (IC5) and the sixth (IC6) independent components were observed to dominantly explain the spectral regions that influenced clustering of diseased and control scores, respectively. Four independent components were instrumental for clustering scores in grade 2 dataset. Analysis shows the fifth (IC5) and seventh (IC7) independent components greatly influenced clustering of controls' scores, whereas the ninth (IC9) and tenth (IC10) independent component greatly influenced clustering of diseased scores. For grade 3 dataset, the eighth (IC8) and ninth (IC9) independent components explains clustering of controls' scores whereas the tenth (IC10) predominantly influenced clustering of diseased scores.

To better understand the biochemical alterations responsible for clustering of control and diseased scores, the spectral regions (both positive and negative bands) associated with the observed ICs were analyzed (Figures 5.23 and 5.24). For clarity, the tentative biochemical assignments were detailed in Table 5.10. Analysis of loading vectors showed aromatic acids proteins were a major factor in clustering of both control and diseased samples. For instance, it can be observed in Figure 5.23 (a)-(b) that biochemical changes due to proteins (596 cm^{-1} , 620 cm^{-1} , 850 cm^{-1} , 858 cm^{-1} , 865 cm^{-1} , 1002 cm^{-1}) were a predominant factor in control samples whereas biochemical changes due to proteins (592 cm^{-1} , 631 cm^{-1} , 855 cm^{-1} , 1005 cm^{-1}), nuclei acids (1160 cm^{-1}) and lipids (1349 cm^{-1}) played a key role during early breast cancer progression. If we consider Figure 5.23 (c-f), it is observed that changes due to proteins (1002 cm^{-1} , 1006 cm^{-1} , 1247 cm^{-1} , 1349 cm^{-1} , 1628 cm^{-1}), lipids (1349 cm^{-1}), nuclei acids (1359 cm^{-1}) were predominant in control samples whereas changes due to proteins (592 cm^{-1} , 630 cm^{-1} , 858 cm^{-1} , 871 cm^{-1} , 1005

cm⁻¹, 1624 cm⁻¹) dominated cancer progression, a trend that was replicated during late breast malignancy (Figure 5.24 (a-c)).

Comparison of the results for PLS-DA and ICA followed by PLS-DA on spectral matrices of selected spectral markers suggests that ICA-PLS-DA performed better in revealing majority of spectral markers that were responsible for discriminating control from diseased samples (Table 5.16). In contrast to other dimensional reduction algorithms such as PCA, ICA identifies non-Gaussian components which are modelled as a linear combination of the biological features (Yao *et al.*, 2012). These components are statistically independent, i.e. there is no overlapping information between the components, a characteristic property that enabled unmasking all spectral markers responsible for samples discrimination, which could not be understood solely by use of PLS-DA. The diagnostic results of using independent components followed by PLS-DA in discrimination of control and diseased scores are provided in Table 5.17.

Table 5.16 Comparison of chemometric results on subtle spectral markers using PLS-DA and ICA-PLS-DA techniques

| PLS-DA | | | | | | ICA-PLS-DA | | | | | |
|-------------------------------------|------|---------|------|---------|------|-------------------------------------|------|---------|------|---------|------|
| Spectral markers / cm ⁻¹ | | | | | | Spectral markers / cm ⁻¹ | | | | | |
| Grade 1 | | Grade 2 | | Grade 3 | | Grade 1 | | Grade 2 | | Grade 3 | |
| CTR* | DIS* | CTR* | DIS* | CTR* | DIS* | CTR* | DIS* | CTR* | DIS* | CTR* | DIS* |
| 593 | 1357 | 594 | 1626 | 1248 | 594 | 596 | 592 | 1002 | 592 | 628 | 589 |
| 631 | 1626 | 632 | 1634 | 1349 | 630 | 620 | 631 | 1006 | 630 | 1245 | 594 |
| | 1630 | | | | 856 | 850 | 855 | 1247 | 858 | 1250 | 616 |
| | | | | | 868 | 858 | 1005 | 1349 | 871 | 1347 | 630 |
| | | | | | 1626 | 865 | 1160 | 1359 | 1005 | 1630 | 848 |
| | | | | | 1631 | 1002 | 1349 | 1628 | 1624 | | 856 |
| | | | | | 1634 | | | | | | 868 |

* The CTR and DIS identify the control and diseased status of the whole blood samples

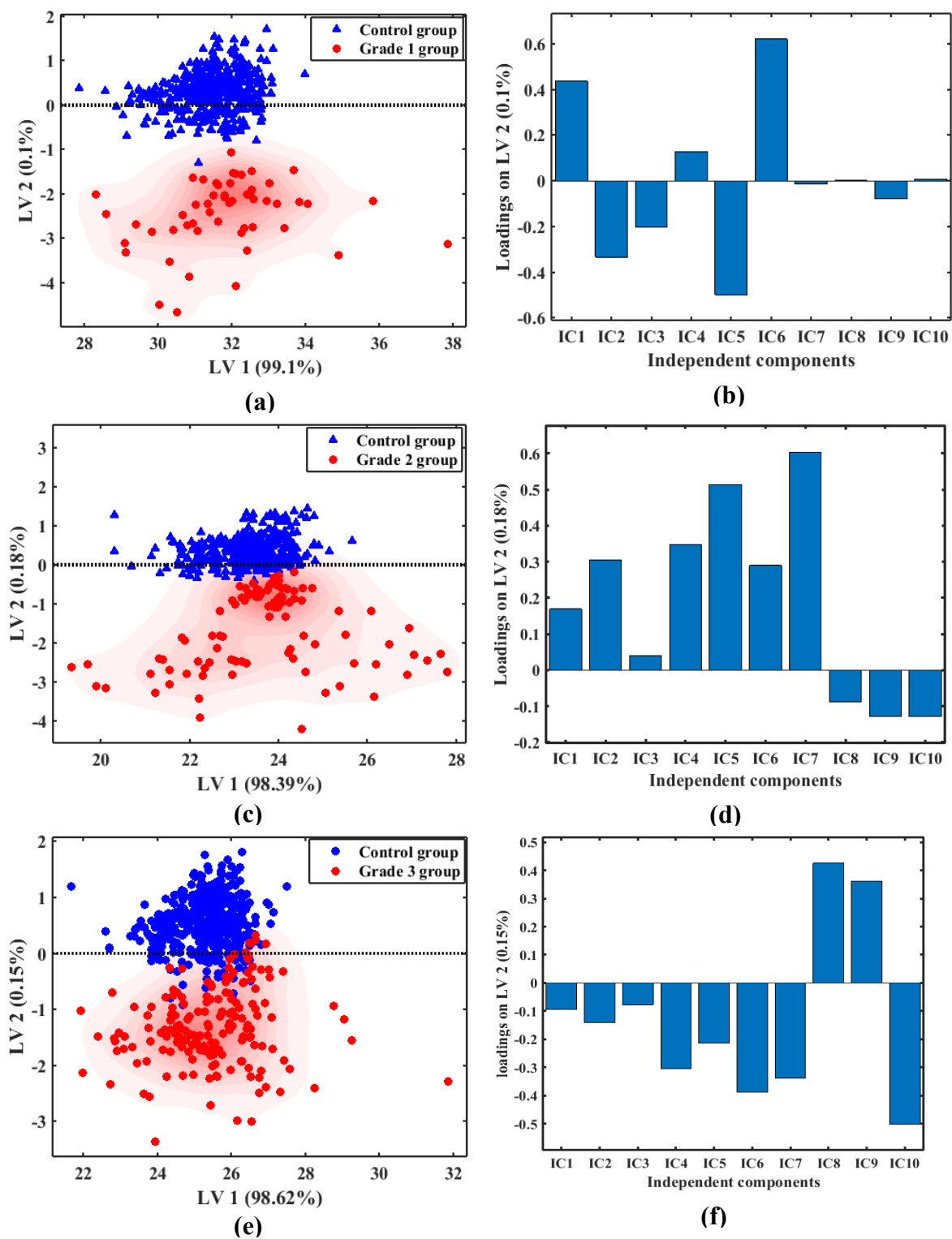


Figure 5.22 The ICA-PLS-DA scatter plots for Raman spectra of blood samples from healthy volunteers and (a) grade 1, (c) grade 2, and (e) grade 3 breast cancer patients. The independent components associated with respective loadings are shown in parts (b), (d), and (f), respectively.

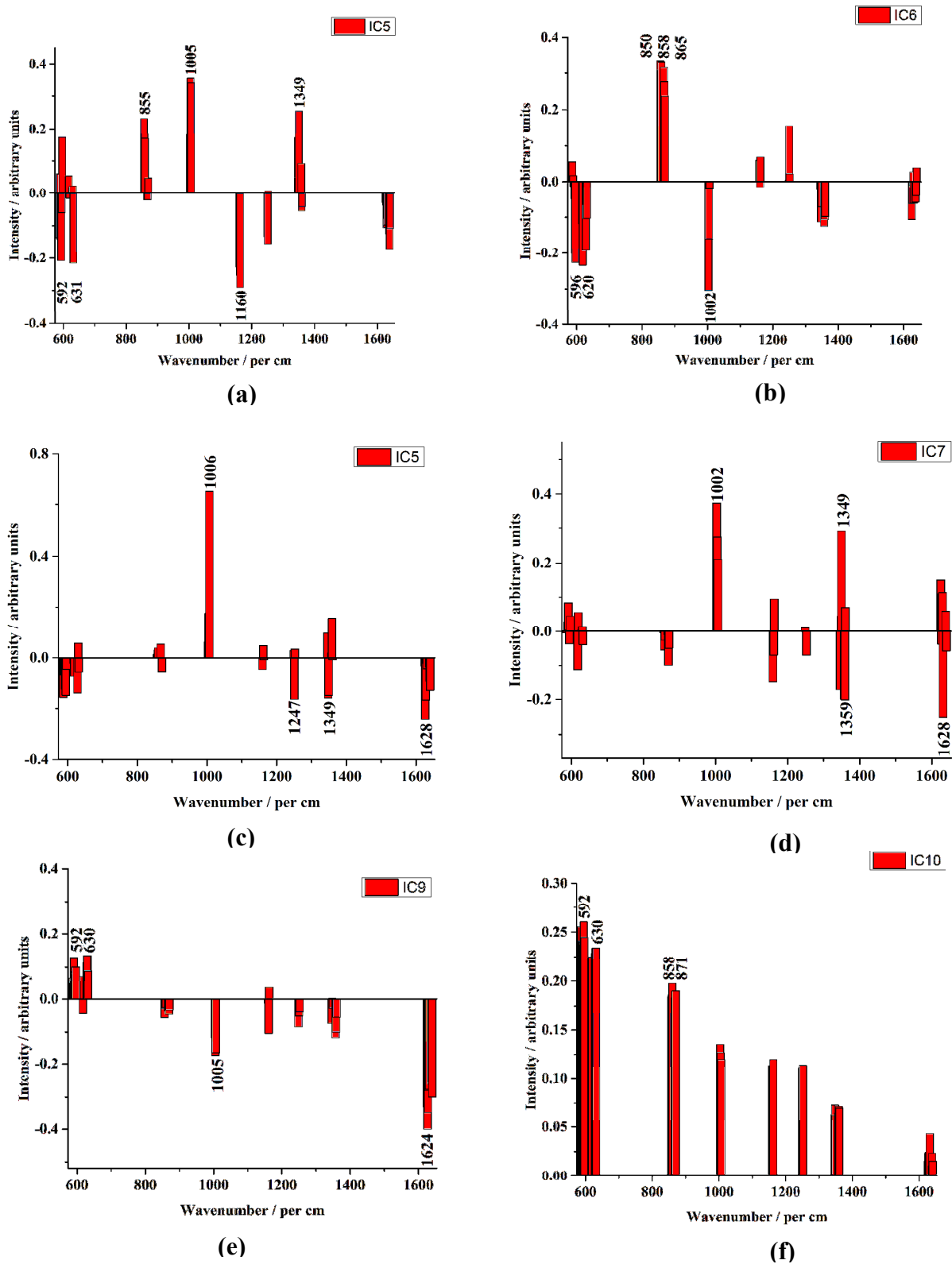
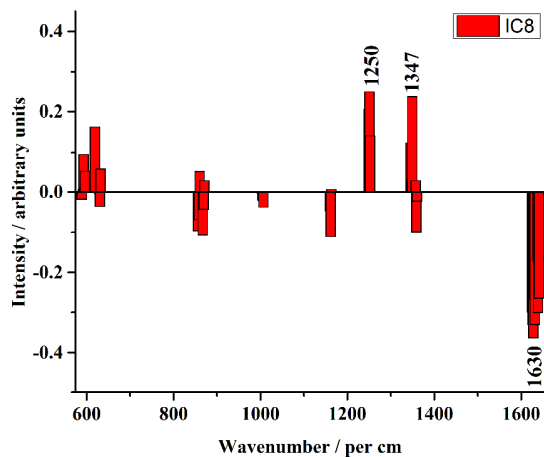
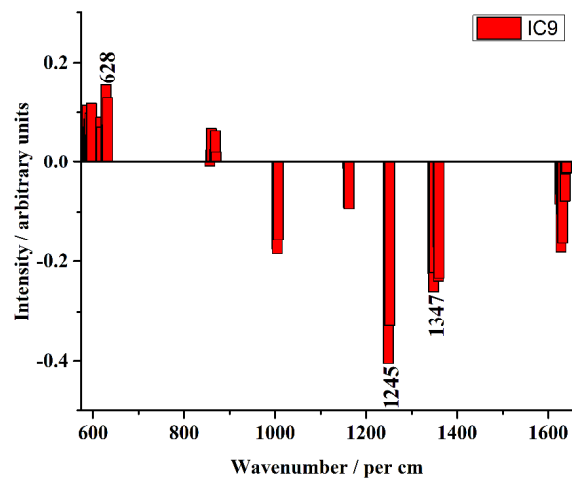


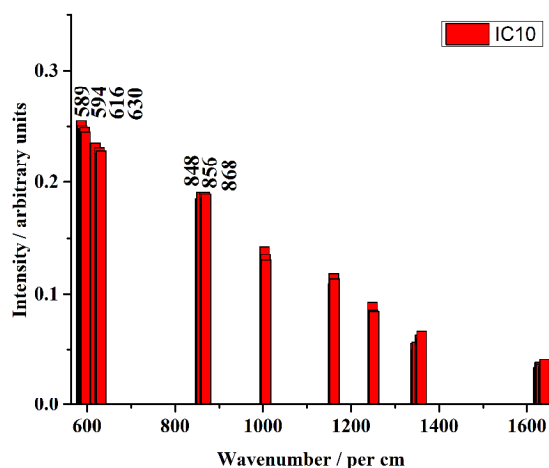
Figure 5.23 The spectral markers for independent components of Raman spectra of blood samples from healthy volunteers and (a), (b) grade 1, and (c), (f) grade 2 breast cancer patients.



(a)



(b)



(c)

Figure 5.24 The spectral markers for (a-c) independent components of Raman spectra of blood samples from healthy volunteers and grade 3 breast cancer patients.

Table 5.17 Diagnostic results of ICA followed by PLS-DA on the Raman spectra of blood from healthy volunteers (controls) and breast cancer patients

| Disease status | Diagnosis | Cases | | | Total | Accuracy | Sensitivity | Specificity |
|----------------|---------------|---------------|----------|--------------|-------|----------|-------------|-------------|
| | | Breast cancer | Controls | Not assigned | | | | |
| Grade-1 | Breast cancer | 54 | 2 | 0 | 56 | 100% | 100% | 100% |
| | Controls | 1 | 424 | 3 | 428 | | | |
| Grade-2 | Breast cancer | 94 | 2 | 1 | 97 | 98% | 100% | 99% |
| | Controls | 1 | 414 | 7 | 422 | | | |
| Grade-3 | Breast cancer | 164 | 12 | 6 | 192 | 93% | 96% | 95% |
| | Controls | 16 | 415 | 3 | 434 | | | |

It is observed that the overall diagnostic accuracies decreased with malignancy, meaning underlying complexity of biochemical alteration in healthy and diseased samples increased with malignancy. This agrees with achieved sensitivity values where it was highest in the first two stages of malignancy, but lowest in late malignancy, implying significant differences in the underlying biochemical changes between blood samples of malignant and control patients. If we consider sensitivity parameters, it can be observed that there was better diagnosis of breast cancer using ICA-PLS-DA analysis (Table 5.17) when compared to similar analysis using PLS-DA only (Table 5.14). For instance, on performing PLS-DA analysis, the sensitivity parameters for grade 2 and grade 3 datasets analysis were 98% and 89%, respectively (Table 5.14) whereas analysis of grade 2 and grade 3 datasets using ICA followed by PLS-DA yielded sensitivities of 100% and 96%, respectively.

The results obtained demonstrates that ICA has capability of producing basis vectors that are statistically independent, and not just linearly decorrelated as it happens with PCA. For that reason, it provides a more powerful data representation and can therefore be used as a discriminate analysis criterion for enhancing PCA. A slight advantage of PCA over ICA is that the resulting vectors are sorted by their importance. It should also be noted that although we can decompose the data in several components with ICA, the ICA algorithm will not tell which one of them is the most important. Therefore, adding a supervised constraint on the ICA for discrimination purpose appears necessary to increase weight of underlying spectral features in the classification, which explains why PLS-DA was included for classification.

In the present study, the ICA analysis was extended to include MDS as a potential non-linear dimensional reduction algorithm in blood Raman datasets. Independent components were analyzed using Minkowski MDS metrics (Weinberg, 1991) and the resultant matrices of coordinates subjected to PLS-DA. The score plots and diagnostic performances are shown in Figure 5.25 and Table 5.18, respectively.

Table 5.18 Diagnostic results of ICA followed by MDS and PLS-DA on the Raman spectra of blood from healthy volunteers (controls) and breast cancer patients

| Disease status | Diagnosis | Cases | | | Total | Accuracy | Sensitivity | Specificity |
|----------------|---------------|---------------|----------|--------------|-------|----------|-------------|-------------|
| | | Breast cancer | Controls | Not assigned | | | | |
| Grade-1 | Breast cancer | 56 | 0 | 0 | 56 | 100% | 100% | 100% |
| | Controls | 1 | 423 | 4 | 428 | | | |
| Grade-2 | Breast cancer | 94 | 0 | 3 | 97 | 100% | 100% | 99% |
| | Controls | 1 | 419 | 3 | 422 | | | |
| Grade-3 | Breast cancer | 187 | 5 | 0 | 192 | 96% | 97% | 96% |
| | Controls | 17 | 415 | 2 | 434 | | | |

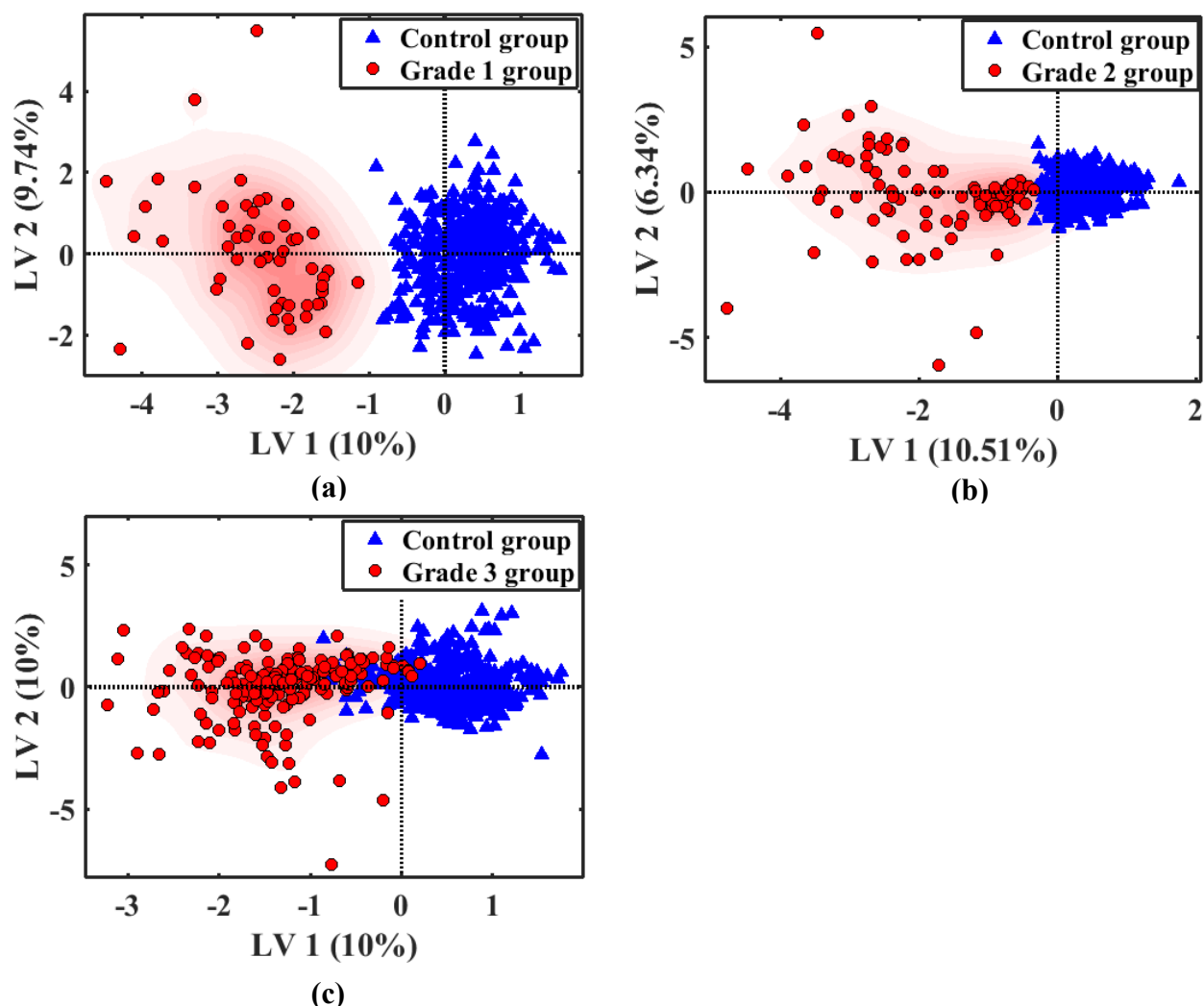


Figure 5.25 The ICA followed by MDS-PLS-DA scatter plots of Raman spectra of blood samples from healthy volunteers and (a) grade 1, (b) grade 2, and (c) grade 3 breast cancers patients.

Comparison of diagnostic results summarized in Table 5.17 and Table 5.18 suggest that inclusion of multidimensional scaling prior to PLS-DA on spectral datasets of blood samples yielded better performance in terms of accuracy and sensitivity. In particular, it marginally yielded a better diagnosis of late (grade 3) malignancy at sensitivity of 97% (Table 5.18) when compared to sensitivity of 96% achieved by ICA followed by PLS-DA (Table 5.17). The better performance of MDS can be attributed to its strength in mapping all pairwise distances between data points into small dimensional Euclidean domains (Aflalo *et al.*, 2013), while preserving the intrinsic information of pairwise dissimilarities between objects (Liu *et al.*, 2019).

5.2.1.5 Multivariate exploratory analysis of Support Vector Machine (SVM) and Backpropagation neural network (BPNN) for breast cancer diagnostics in blood

Two diagnostic models were chosen; the support vector machine (SVM) and backpropagation neural networks (BPNN). The use of SVM was motivated by its great performance in handling both linear and high non-linear data (Singla *et al.*, 2011), whereas BPNN was chosen due to flexibility of optimizing its architecture during feature selection (Bisgin *et al.*, 2018). The radial basis function (RBF) has been reported to outperform other kernel functions in nonlinear classification (Bisgin *et al.*, 2018), and was therefore chosen as one of kernel functions to solve the SVM classifiers in our study. The classifiers used the Raman measurements taken at the Raman shifts of 589, 594, 630, 858, 868, 1005, 1160, 1250, 1347, 1358, 1626, 1630, and 1638 cm^{-1} . These 13 spectral bands explained the trace spectral markers determined by the use of intermediate- and high-order principal components as described in section 5.2.1.2. PCA was used to achieve minimal redundancy during feature selection (Crow *et al.*, 2005), which allowed better understanding of support vectors features that were more relevant to scores discrimination. About 5-10 principal components were selected for SVM cross-validation procedure, where the $k(= 10)$ -fold cross-validation testing procedure was assessed in terms of classifier accuracy, sensitivity and specificity. Figure 5.26 shows the resultant SVM scatter plots for (a, b) grade 1 breast cancer, (c, d) grade 2 breast cancer, and (e, f) grade 3 breast cancer. Figure 5.26 (a), (c) and (e) are scatter plots of models based on linear SVM kernel function, whereas (b), (d) and (f) are scatter plots of models based on radial basis function (RBF) SVM kernel function. Details of linear and RBF-SVM functions and respective classifier performance parameters are provided in Table 5.19 and 5.20, respectively.

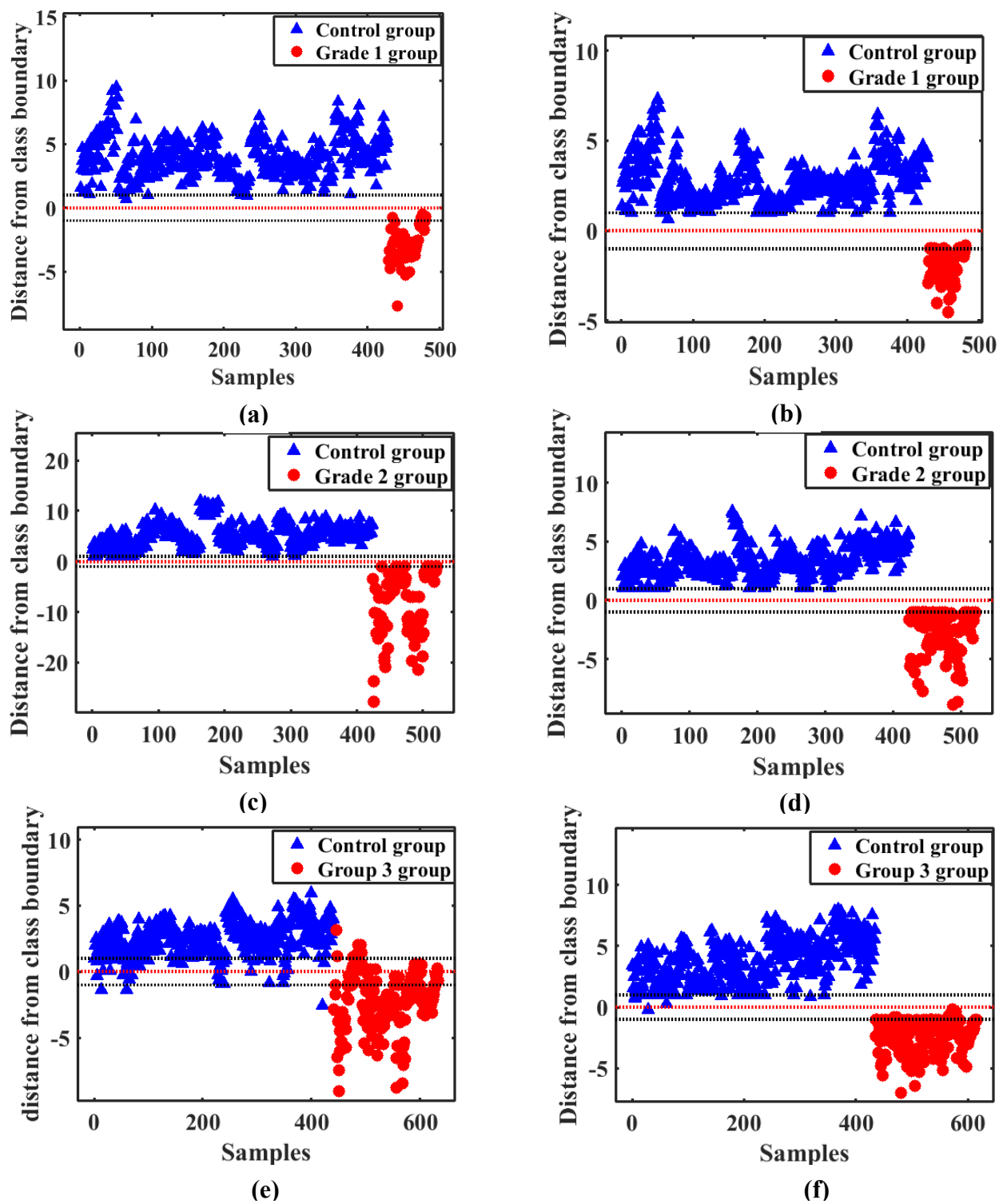


Figure 5.26 The SVM scatter plots for breast cancer detection of (a, b) grade 1 ($n = 26$), (c, d) grade 2 ($n = 30$), and (e, f) grade 3 breast cancer ($n = 33$) spectral datasets. Parts (a), (c) and (e): linear kernel function scatter plots; (b), (d) and (f): radial basis function (RBF) scatter plots.

Table 5.19 SVM models characteristics for diagnostic analysis on the Raman spectra of whole blood from healthy volunteers (controls) and breast cancer patients

| SVM optimal characteristic | | | | | |
|----------------------------|----------|------|-----|-------------------|-----------------|
| Disease status | Function | Cost | PCs | Kernel parameters | Support vectors |
| Grade-1 | Kernel | 100 | 5 | - | 14 |
| | RBF | 100 | 5 | 0.8 | 19 |
| Grade-2 | Kernel | 100 | 10 | - | 9 |
| | RBF | 100 | 10 | 0.8 | 27 |
| Grade-3 | Kernel | 100 | 10 | - | 118 |
| | RBF | 100 | 10 | 0.57 | 67 |

Table 5.20 Diagnostic results of linear-SVM and RBF-SVM models on the Raman spectra of whole blood from healthy volunteers (controls) and breast cancer patients

| Cases | | | | | | | | |
|----------------|----------|---------------|---------------|----------|-------|--------------|-----------------|-----------------|
| Disease status | Function | Diagnosis | Breast cancer | Controls | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
| Grade-1 | Linear | Breast cancer | 53 | 0 | 53 | 100 | 100 | 100 |
| | | Controls | 0 | 428 | 428 | | | |
| | RBF | Breast cancer | 52 | 1 | 53 | 100 | 98 | 100 |
| | | Controls | 0 | 428 | 428 | | | |
| Grade-2 | Linear | Breast cancer | 98 | 0 | 98 | 100 | 100 | 100 |
| | | Controls | 0 | 423 | 423 | | | |
| | RBF | Breast cancer | 98 | 0 | 98 | 100 | 98 | 100 |
| | | Controls | 0 | 423 | 423 | | | |
| Grade-3 | Linear | Breast cancer | 163 | 30 | 193 | 93 | 84 | 96 |
| | | Controls | 17 | 423 | 440 | | | |
| | RBF | Breast cancer | 183 | 10 | 193 | 98 | 95 | 99 |
| | | Controls | 5 | 418 | 423 | | | |

For BPNN, learning was carried out by regulating the weights using error feedback from the training samples in order to bring the network prediction of the correct outputs for the training samples closer to the true values. For blood Raman datasets, it was observed that 2 layers, neurons=25, learning rate=0.001, alpha=2, and iterations=1000 training weights were useful for optimal classification into control and diseased scores (groups). Figure 5.27 shows scores classification of Raman spectra from blood of control and diseased patients based on BPNN diagnostic model. The accuracy, sensitivity and specificity classification parameters of BPNN diagnostic model are summarized in Table 5.21.

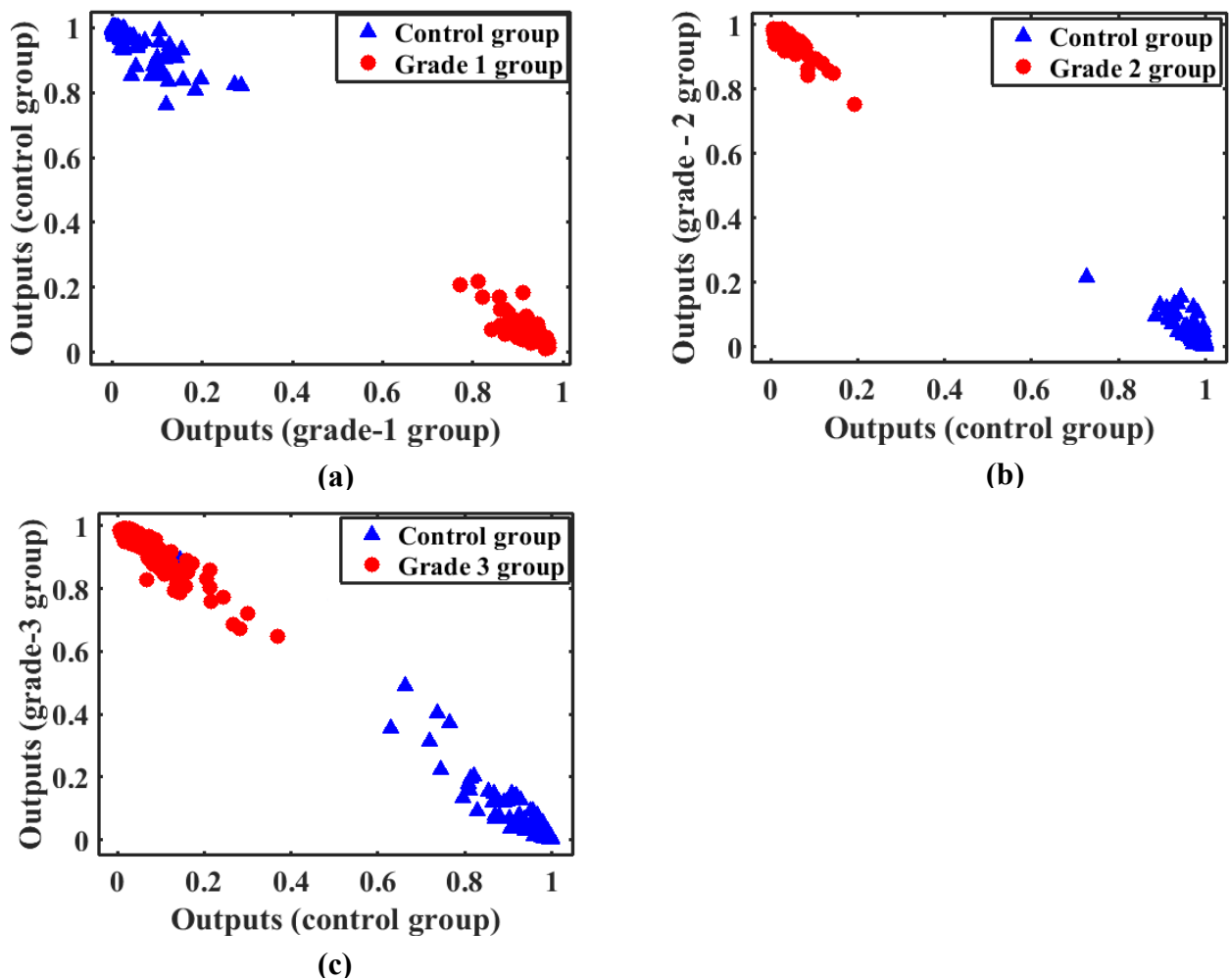


Figure 5.27 The scatter plots of BPNN diagnostics model for detection of (a) grade 1, (b) grade 2, and (c) grade 3 breast cancer.

Table 5.21 Diagnostic results of BPNN diagnostic model on the Raman spectra of whole blood from healthy (controls) and breast cancer patients

| Disease status | Diagnosis | Cases | | | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|----------------|---------------|---------------|----------|--------------|-------|--------------|-----------------|-----------------|
| | | Breast cancer | Controls | Not assigned | | | | |
| Grade-1 | Breast cancer | 46 | 1 | 6 | 53 | 100 | 98 | 100 |
| | Controls | 0 | 402 | 26 | 428 | | | |
| Grade-2 | Breast cancer | 93 | 2 | 2 | 97 | 99 | 98 | 98 |
| | Controls | 2 | 419 | 1 | 422 | | | |
| Grade-3 | Breast cancer | 162 | 10 | 10 | 182 | 97 | 95 | 96 |
| | Controls | 7 | 424 | 3 | 434 | | | |

It can be seen in Table 5.19 that both SVM models achieved optimal performance with same cost function (C), which suggests both models yielded similar classification error terms (Bouzalmat *et al.*, 2014). However, greater number of principal components (=10) and support vectors (27-120) were needed for optimal analysis of grade 2 and grade 3 datasets in comparison to number of principal components (=5) and support vectors (14-20) required for analysis of grade 1 dataset. This suggests that grade 2 and grade 3 breast malignancy matrices greatly suffered from problems of high dimensionality and collinearity which necessitated higher number of principal components to account for greater amount of variance in the datasets (Björklund, 2019). Consequently, a relatively greater number of support vectors were needed to optimally define a hyperplane for maximizing margins between the two classes (controls versus the diseased scores) (Martins *et al.*, 2009).

Table 5.20 shows that diagnostic accuracies decreased with level of malignancy where it is observed that both SVM diagnostic models performed very well in diagnosing grade 1 and 2 malignancy with overall classification accuracies at 100% with sensitivities and specificities in the range of 98% -100%. In terms of classification accuracy (Table 5.20), RBF kernel function model performed better than linear kernel function model in diagnosing late (grade 3) malignancy, where the overall classification accuracies of linear and RBF models were 93% and 98%, respectively whereas the sensitivities were 84% and 95%, respectively. This suggests that the linear separable characteristic nature of spectral datasets decreased with malignancy. It should be

noted that, in comparison to linear kernels (parametric functions), RBF is a squared exponential function (non-parametric function) and can therefore be viewed as powerful as an infinite order polynomial kernel (Chen *et al.*, 2015). Compared to parametric model (e.g. linear kernel functions), the complexity of the nonparametric model (e.g. RBF kernel functions) is potentially infinite (Chen *et al.*, 2015), suggesting its complexity can grow with the data and can therefore represent more and more complex relationships (Sajda, 2006). So asymptotically, assuming you have unlimited data and very weak assumptions about the problem, a nonparametric method is generally better and expected to have better performance in samples discrimination (Mika *et al.*, 2002).

Table 5.21 shows the overall diagnostic accuracies of the BPNN diagnostic model was generally excellent (> 95%). However, sensitivity of BPNN diagnostic model decreased with malignancy which can be attributed to the complex nature of biochemical alterations involved during progression of malignancy. Nevertheless, comparison of performances of SVM and BPNN diagnostic models on Raman spectra of whole blood samples suggests they generally achieved same performance.

Having understood the performance of SVM and BPNN models on Raman spectra, the SVM and BPNN prediction diagnostic models were designed with aim of predicting the disease status of other 'unknown' liquid biopsy samples (test set), randomly chosen and Raman measurements taken using the same configuration. In this context, the 'unknown' liquid biopsy sample (test set) refers to spectra collected from independently selected blood samples from the healthy (controls) and diseased (breast cancer) patients. As expected, the prediction models were performed on spectral measurements taken at the Raman shifts of 589, 594, 630, 858, 868, 1005, 1160, 1250, 1347, 1358, 1626, 1630, and 1638 cm^{-1} . When performing SVM prediction, a voting strategy was employed in determining the 'selected class' (Happillon *et al.*, 2015), i.e., samples were clustered to classes chosen by majority of SVMs. For linear SVM predictor model, the selected optimal prediction parameters included: cost = 100, number of principal components = 5-10, number of support vectors = 20-52, based on 10-fold cross-validation method. Similarly, it was observed the prediction of healthy (control) and diseased samples by RBF SVM model could be optimally achieved by adopting the following tuning parameters: kernel parameter = 0.57-0.8, cost = 100, support vectors = 20-44, number of principal components = 5-10, based on 10-fold cross-validation method. The prediction abilities of the SVM classifiers by linear and RBF kernel

functions are shown in Figure 5.28 (a, c, e) and (b, d, f), respectively. The respective confusion matrices for the tests data, using k -fold validation are provided in Table 5.22.

Table 5.22. Diagnostic results of linear and RBF SVM predictor models on the Raman spectra of whole blood from healthy (controls) and breast cancer patients

| | | Cases | | | | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|----------------|----------|---------------|---------------|----------|-----|-------|--------------|-----------------|-----------------|
| Disease status | Function | Diagnosis | Breast cancer | Controls | | | | | |
| Grade-1 | Linear | Breast cancer | 18 | 0 | 18 | 98 | 100 | 98 | |
| | | Controls | 3 | 145 | 148 | | | | |
| | RBF | Breast cancer | 18 | 0 | 18 | 98 | 100 | 98 | |
| | | Controls | 3 | 145 | 148 | | | | |
| Grade-2 | Linear | Breast cancer | 32 | 3 | 35 | 96 | 91 | 97 | |
| | | Controls | 5 | 143 | 148 | | | | |
| | RBF | Breast cancer | 33 | 2 | 35 | 95 | 94 | 95 | |
| | | Controls | 7 | 141 | 148 | | | | |
| Grade-3 | Linear | Breast cancer | 50 | 14 | 64 | 93 | 78 | 99 | |
| | | Controls | 1 | 151 | 156 | | | | |
| | RBF | Breast cancer | 57 | 7 | 64 | 94 | 89 | 97 | |
| | | Controls | 5 | 147 | 152 | | | | |

For BPNN predictor model, the training weights of the number of layers = 2, neurons=25, learning rate=0.0001, alpha = 2, and the number of iterations=1000 were observed to deliver optimal prediction and classification of diseased and control samples. The scores prediction are shown in Figure 5.29 and prediction accuracies are summarized in Table 5.23. Based on the previous performance of SVM and BPNN diagnostic models during validation step (Table 5.20), it can be noted that the diagnostic accuracies of prediction models were >90% but decreased with malignancy progression. Furthermore, the sensitivity decreased with malignancy with late malignancy demonstrating sensitivity <90%, suggesting that there were additional spectral features that were insignificant for cancer discrimination.

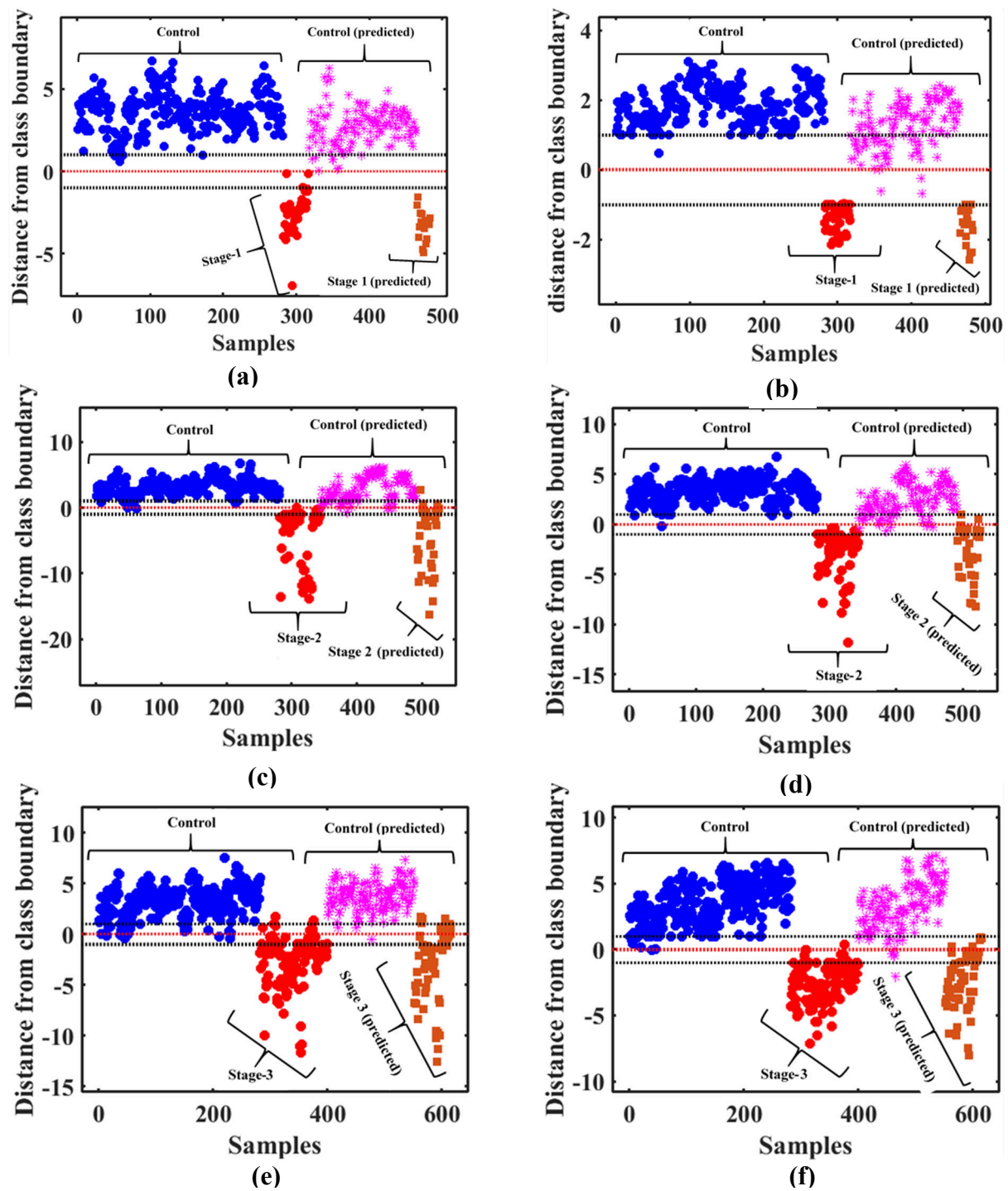


Figure 5.28 SVM prediction models for breast cancer detection of (a, b) grade 1 cancer, (c, d) grade 2 cancer, and (e, f) grade 3 cancer, based on linear (a, c, e) and RBF (b, d, f) kernel functions.

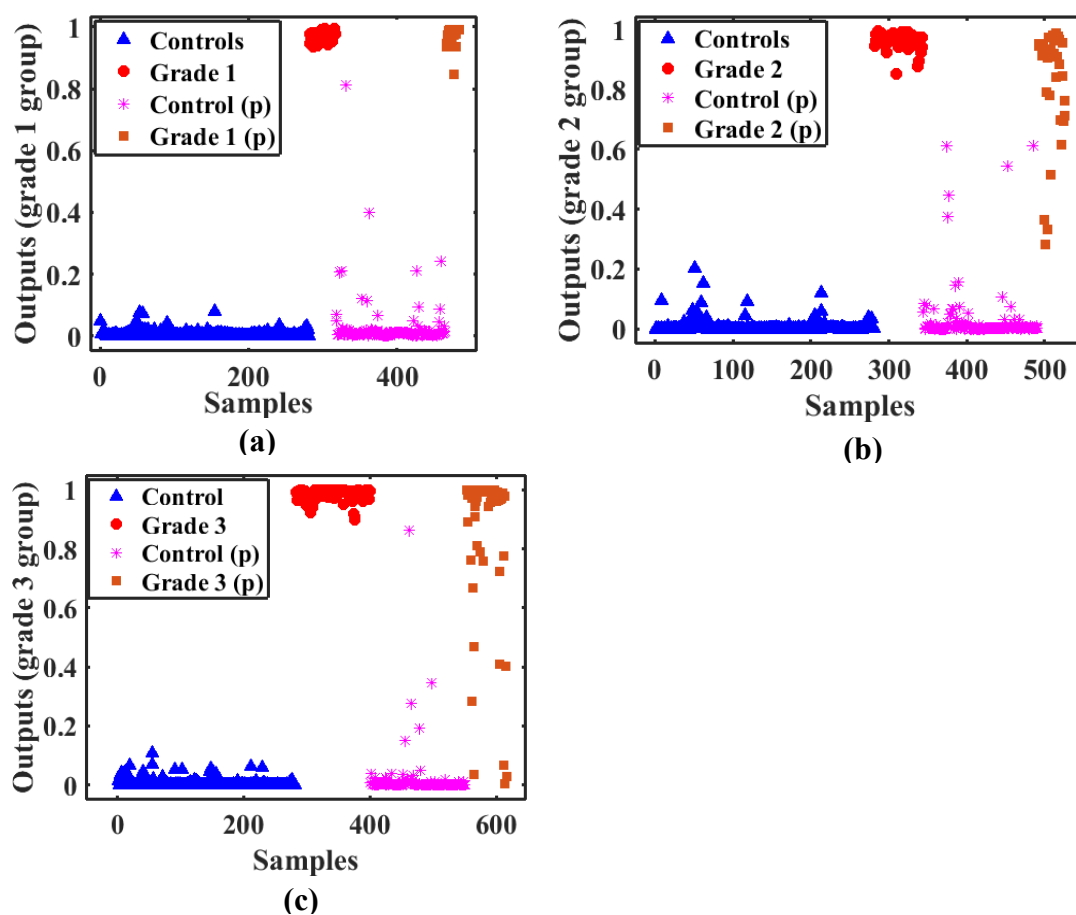


Figure 5.29 BPNN predictor models for (a) grade 1 cancer, (b) grade 2 cancer and (c) grade 3 cancer. For clarity, letter p explains the predicted scores.

Table 5.23 Diagnostic results of BPNN predictor model on the Raman spectra of whole blood from healthy volunteers (controls) and breast cancer patients

| | | Cases | | | | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|----------------|---------------|---------------|----------|--------------|-----|-------|--------------|-----------------|-----------------|
| Disease status | Diagnosis | Breast cancer | Controls | Not assigned | | | | | |
| Grade-1 | Breast cancer | 18 | 0 | 0 | 18 | 99 | 100 | 99 | |
| | Controls | 1 | 146 | 1 | 148 | | | | |
| Grade-2 | Breast cancer | 32 | 1 | 2 | 35 | 98 | 97 | 98 | |
| | Controls | 3 | 144 | 1 | 147 | | | | |
| Grade-3 | Breast cancer | 57 | 5 | 2 | 62 | 97 | 92 | 99 | |
| | Controls | 1 | 151 | 0 | 152 | | | | |

By comparison, the SVM predictor model based on RBF kernel function performed better in diagnosing stage 2 and stage 3 malignancy, which agrees with the performance of nonparametric models (Chen *et al.*, 2015; Sajda, 2006). If we consider Table 5.23, the diagnostic performance parameters suggest that prediction accuracy of BPNN model decreased with malignancy, though prediction accuracies were general better than prediction accuracies achieved by SVM linear models, which confirms its potential strength in samples prediction. Nevertheless, in agreement with observations made in Table 5.21 the sensitivity parameters decreased with malignancy, though specificity relatively remained $\approx 100\%$. The comparison of SVM and BPNN predictor models shows that BPNN outperformed SVM for the score prediction using the present data set (Table 5.22, Table 5.23). This could be due to better parameter selection or the diverse and non-linear nature of the data set or both.

The results obtained strengths the view that SVM and BPNN algorithms are suitable for handling high complexity spectral datasets (Singla *et al.*, 2011). Ideally, SVM incorporates the ability to discriminate non-linearly separable classes that are not characteristic of other multivariate analysis such as the LDA, and is therefore suitable for classifying larger sample sizes whose spectra data may possess non-linear characteristics (Luo *et al.*, 2008). It should however be noted that, for SVMs, the optimal generalization performance is achieved with high dimensionality data and / or dataset with a low training samples to input dimensionality ratio (Belousov *et al.*, 2002). As we have observed, despite the small sample sizes (and therefore dataset matrices), the lowest diagnostic accuracies was $> 90\%$, and sensitivity ranged between 80 to 100%. This confirms that SVMs are suited to dealing with high dimensional spectral data, suggesting they have capability of handling a high degree of collinearity in the datasets. SVM employs kernel tricks and maximal margin concepts to perform better in non-linear and high-dimensional tasks. However, even a powerful model (e.g. SVM) benefit from the proper feature extraction / transformation techniques. The improved performance of the SVM diagnostic model can therefore be attributed to inclusion of PCA as a preprocessing step before cross-validation of datasets. Incorporation of PCA into machine learning is useful for the classification of high dimensional data, since it can alleviate potential problems such as high dimensionality and collinearity that are associated with spectral data (Björklund, 2019). In this study, PCA performed initial transformation of the dataset into a smaller set of PCs, and machine learning classification errors were noted to reduce with increased number of principal components.

On the other hand, artificial neural networks have characteristic of handling multiclass problems and data non-linearity very well (Wang *et al.*, 2014), which generally explain why they performed better in diagnosing and predicting late malignancy (stage 3). It could also be due to the fact that the BPNN converges on a global minimum and allows a better tolerance to the noise (deviation from the pattern that often inherently associated with the original spectra) therefore might be slightly more robust for a large set of features. This could have been facilitated by the careful choice of layer and neurons per layer that minimized challenges associated with model overfitting, greater training time and vanishing / exploding gradients problem.

5.2.2 Raman spectroscopy characterization of saliva for breast cancer diagnosis

5.2.2.1 Analysis of prominent biochemical alterations in saliva spectra

Figure 5.30 (a) shows the optical photomicrograph of a saliva sample. The image depicts crystalline structures conjoined with each other, in a tree or fern-like form from the center of the drop. Other studies (Gonchukov *et al.*, 2012), have suggested that the changes of morphological picture of dried oral saliva fluid are an indication of deviation in quantitative and quality molecular composition. Figure 5.30 (b) shows prominent spectral regions in saliva of healthy volunteers (normal) and breast cancer patients, in the 550-1800 cm^{-1} region. As observed, common mean (\pm standard deviations) Raman band alterations occurred at $623 \pm 1.73 \text{ cm}^{-1}$, 745 cm^{-1} , $821 \pm 1.73 \text{ cm}^{-1}$, 939 cm^{-1} , $1000 \pm 2.02 \text{ cm}^{-1}$, $1125 \pm 0.86 \text{ cm}^{-1}$, $1444 \pm 1.44 \text{ cm}^{-1}$, $1603 \pm 1.44 \text{ cm}^{-1}$, and 1661 cm^{-1} . Besides, there were other bands associated with diseased samples at 715 cm^{-1} , 1227 cm^{-1} , 1499 cm^{-1} , 1558 cm^{-1} and 1637 cm^{-1} . Further, normal samples exhibited heightened band intensity at 1246 cm^{-1} . The respectful biochemical assignments are detailed in Table 5.24.

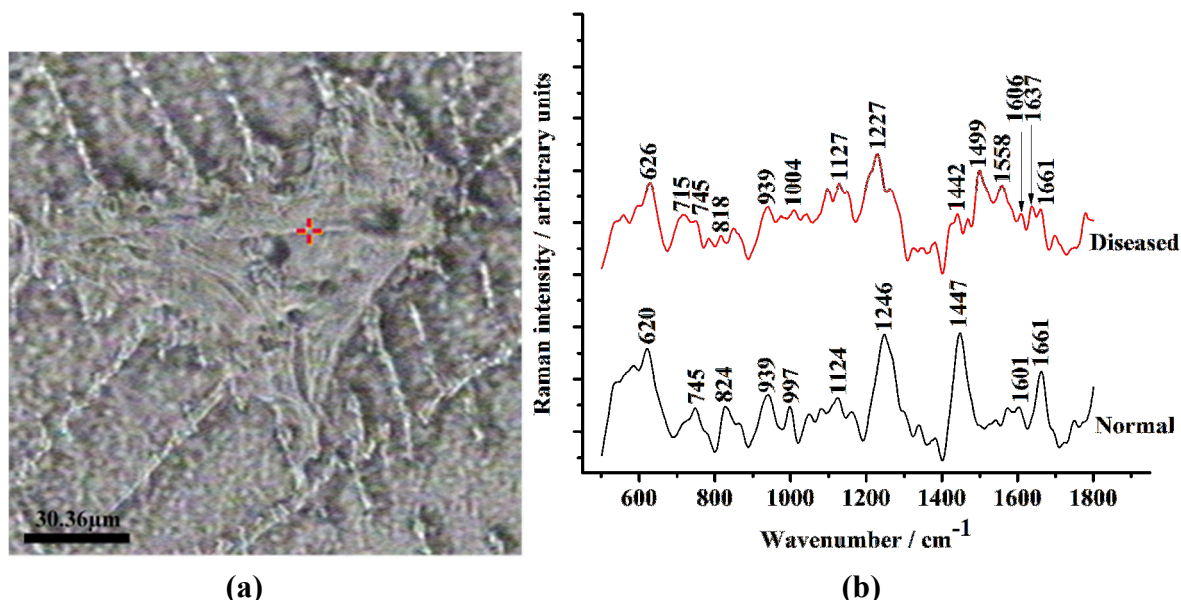


Figure 5.30 (a) Photomicrograph of dried saliva with a laser spot (+) indicated at x50 magnification, and (b) prominent saliva Raman bands amongst the control (normal) and diseased samples.

To better understand differences between the two groups, spectral differences were calculated by subtracting the normalized mean Raman intensity of control samples from the normalized mean Raman intensities of diseased samples (Figure 5.31 (a-d)). At 578 cm^{-1} , 614 cm^{-1} , 664 cm^{-1} , 831 cm^{-1} , 1250 cm^{-1} , 1306 cm^{-1} , 1338 cm^{-1} , 1451 cm^{-1} , and 1665 cm^{-1} spectral regions (Students *t*-test, $p < 0.05$), the peak intensities were greater for the control group than for the breast cancer group, while the spectral regions at 701 cm^{-1} , 796 cm^{-1} , 975 cm^{-1} , 1015 cm^{-1} , 1096 cm^{-1} , 1147 cm^{-1} , 1197 cm^{-1} , 1496 cm^{-1} , 1558 cm^{-1} , 1634 cm^{-1} , and 1705 cm^{-1} were more intense in the saliva of the breast cancer patients (Figure 5.31 (a)). If we consider Figures 5.31 (b-d), it can be observed that the biochemical changes due to lipids (581 cm^{-1} , $607 \pm 1.24\text{ cm}^{-1}$, $939 \pm 1.24\text{ cm}^{-1}$, $1336 \pm 1.02\text{ cm}^{-1}$, $1451 \pm 0.81\text{ cm}^{-1}$, $1661 \pm 0.70\text{ cm}^{-1}$), proteins ($834 \pm 0.47\text{ cm}^{-1}$, $881 \pm 2.62\text{ cm}^{-1}$, $1248 \pm 2.09\text{ cm}^{-1}$, $1336 \pm 1.02\text{ cm}^{-1}$, $1451 \pm 0.81\text{ cm}^{-1}$, $1598 \pm 1.54\text{ cm}^{-1}$, $1661 \pm 0.70\text{ cm}^{-1}$) and nucleic acids ($1336 \pm 1.02\text{ cm}^{-1}$) were consistently greater for the control group than for the breast cancer group, while the changes due to nucleic acids (785 , $1014 \pm 0.70\text{ cm}^{-1}$, $1286 \pm 0.57\text{ cm}^{-1}$, 1495), saccharides ($911 \pm 1.47\text{ cm}^{-1}$, $1145 \pm 0.23\text{ cm}^{-1}$), lipids ($1092 \pm 3.06\text{ cm}^{-1}$, $1371 \pm 0.47\text{ cm}^{-1}$), and proteins (1495 , $1554 \pm 0.86\text{ cm}^{-1}$) were more intense in the saliva of the breast cancer patients when compared to control group. Therefore, it can be concluded that biochemical changes of nucleic acids, proteins, and lipids in saliva can be associated with onset and progression of breast cancer. Detailed biochemical assignments are provided in Table 5.24.

5.2.2.2 Analysis of trace biochemical alterations in saliva spectra

Examination of log scree plots for grade₁, grade₂ and grade₃ spectral datasets from saliva samples of control and breast cancer patients (Figure 5.32 (a-c)) and categorization of PCs based on the cumulative percentage of total variation and the size of variances (Table 5.25) showed that the number of PCs with eigenvalue >1 were 350, 378, and 406, respectively. Further analysis by Kaiser's method (Martinez *et al.*, 2005) suggested that 12, 24 and 17 PCs were potentially useful for further spectral analysis of grade₁, grade₂ and grade₃ spectral datasets, respectively. Further analysis of PCs was determined by help of canonical variable distribution plots (Figure 5.33).

The first (PC 1) and second principal components (PC 2) explained cumulative variances of 78.07%, 78.65%, and 83.87% for grade₁, grade₂ and grade₃ spectral datasets, respectively. PCs 9, 10 and 2 had the largest canonical loading parameters (Figure 5.33) for grade₁, grade₂ and grade₃ spectral datasets respectively, that suggested their potential strength for higher classification accuracies in samples discrimination. To better understand the influence of PCs 9, 10 and 2 in

scores discrimination, the statistical values were calculated (Table 5.26). Examination of Table 5.26 shows that PCs 9, 10 and 2 were the most significant and had the largest effect sizes in their respective datasets. These PCs represented 0.61%, 0.38%, and 10.47% of the total variance in their respective input data.

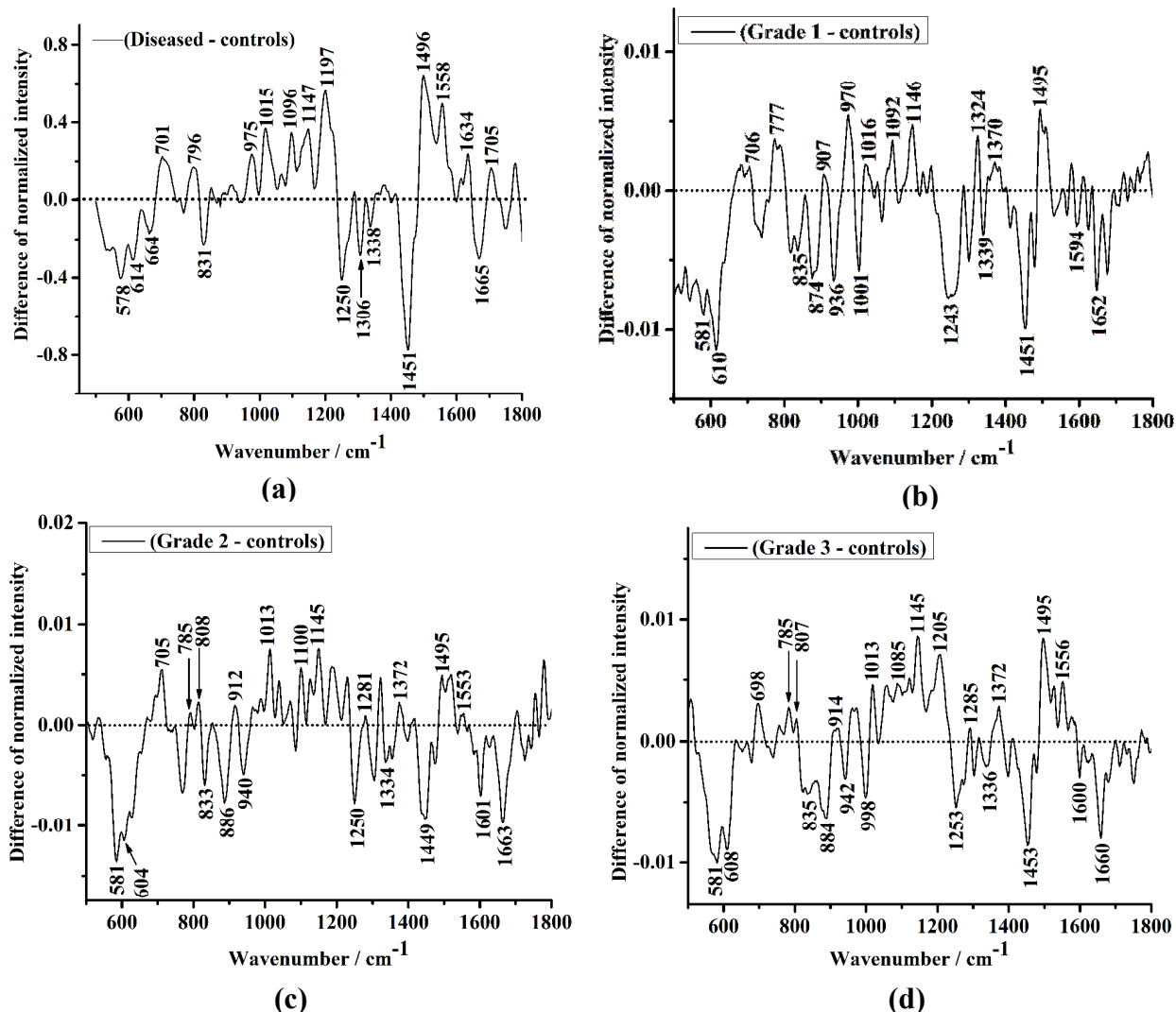


Figure 5.31 Overall spectra differences between Raman spectra of (a) the control ($n = 23$) and all diseased ($n = 20$) samples, (b) the control ($n = 23$) and grade 1 ($n = 3$) samples, (c) the control ($n = 23$) and grade 2 ($n = 7$) samples, and (d) the control ($n = 23$) and grade 3 ($n = 10$) samples.

Table 5.24 Raman band assignments of saliva from healthy volunteers and breast cancer patients

| Raman shift (cm^{-1}) | Functional groups and molecular vibration assignments | References |
|-------------------------------------|---|---|
| 578, 581 | Symmetric stretching vibrations (phosphatidylinositol) | (Rehman <i>et al.</i> , 2013) |
| 614, 607 \pm 1.24 | Cholesterol esters | (Chandra <i>et al.</i> , 2015) |
| 664 | C-C twisting modes of phospholipids and proteins, ring breathing modes of DNA / RNA bases | (Pichardo-Molina <i>et al.</i> , 2007) (Rehman <i>et al.</i> , 2013) |
| 701 | ν (C-S) twisting / stretch modes of cholesterol esters | (Movasaghi <i>et al.</i> , 2007) |
| 785, 796 | DNA: O–P–O stretch of cytosine, uracil, and thymine | (Rehman <i>et al.</i> , 2013) |
| 831, 834 \pm 0.47 | Asymmetric O P O stretching of tyrosine | (Pichardo-Molina <i>et al.</i> , 2007) |
| 881 \pm 2.62 | δ (ring)modes of tryptophan | (Gelder <i>et al.</i> , 2007) |
| 911 \pm 1.47 | C-O-C skeletal mode (glucose) | (Rehman <i>et al.</i> , 2013) |
| 939 \pm 1.24 | Skeletal stretch α | (Pichardo-Molina <i>et al.</i> , 2007) |
| 975 | δ (=CH wagging) of lipids, | (Rehman <i>et al.</i> , 2013) |
| 1015, 1014 \pm 0.70 | ν (C-O) stretch of DNA ribose | (Rehman <i>et al.</i> , 2013) |
| 1096, 1092 \pm 3.06 | O-P-O and C-C stretch (phospholipids) | (Vargas-Obieta <i>et al.</i> , 2016) |
| 1147, 1145 \pm 0.23 | C–O stretching mode of saccharides | (Gelder <i>et al.</i> , 2007) |
| 1197 | Antisymmetric phosphate vibrations (nucleic acids) | (Gelder <i>et al.</i> , 2007) |
| 1250, 1248 \pm 2.09 | β – sheet (Amide III), CH ₂ wagging of glycine and proline | (Chandra <i>et al.</i> , 2015) |
| 1286 \pm 0.57 | Phosphodiester groups in nucleic acids, cytosine | (Rehman <i>et al.</i> , 2013) |
| 1306 | CH ₃ / CH ₂ twisting or bending mode of lipid and collagen, ring breathing modes of DNA / RNA | (Rehman <i>et al.</i> , 2013) |
| 1338, 1336 \pm 1.02 | δ (CH ₃), δ (CH ₃), twisting of proteins, phospholipids helix, ring breathing modes in the DNA bases | (Vargas-Obieta <i>et al.</i> , 2016) |
| 1371 \pm 0.47 | ν_s (CH ₃) stretch of phospholipids | (Chandra <i>et al.</i> , 2015) |
| 1451, 1451 \pm 0.81 | δ (CH ₂), δ (CH ₃), scissoring of lipids, and proteins | (Pichardo-Molina <i>et al.</i> , 2007) |
| 1495, 1496 | C-N stretch / C-N bending (proteins), ring breathing modes of cytosine) | (Rehman <i>et al.</i> , 2013) |
| 1558, 1554 \pm 0.86 | ν (CN) and δ (NH) amide II (protein assignment) | (Gelder <i>et al.</i> , 2007) |

Table 5.24 (continued) Raman band assignments of saliva from healthy volunteers and breast cancer patients

| Raman shift (cm^{-1}) | Functional groups and molecular vibration assignments | References |
|-------------------------------------|--|--------------------------|
| 1598 \pm 1.54 | $\delta(\text{C}=\text{O})$ stretch of amide I | (Rehman et al., 2013) |
| 1634 | $\delta(\text{C}=\text{O})$ stretch of amide I | (Rehman et al., 2013) |
| 1665, 1661 \pm 0.70 | C=O stretch (proteins, lipids), C = C stretching of lipids | (Movasaghi et al., 2007) |
| 1715 | C=O stretch of amide groups, thymine | (Rehman et al., 2013) |

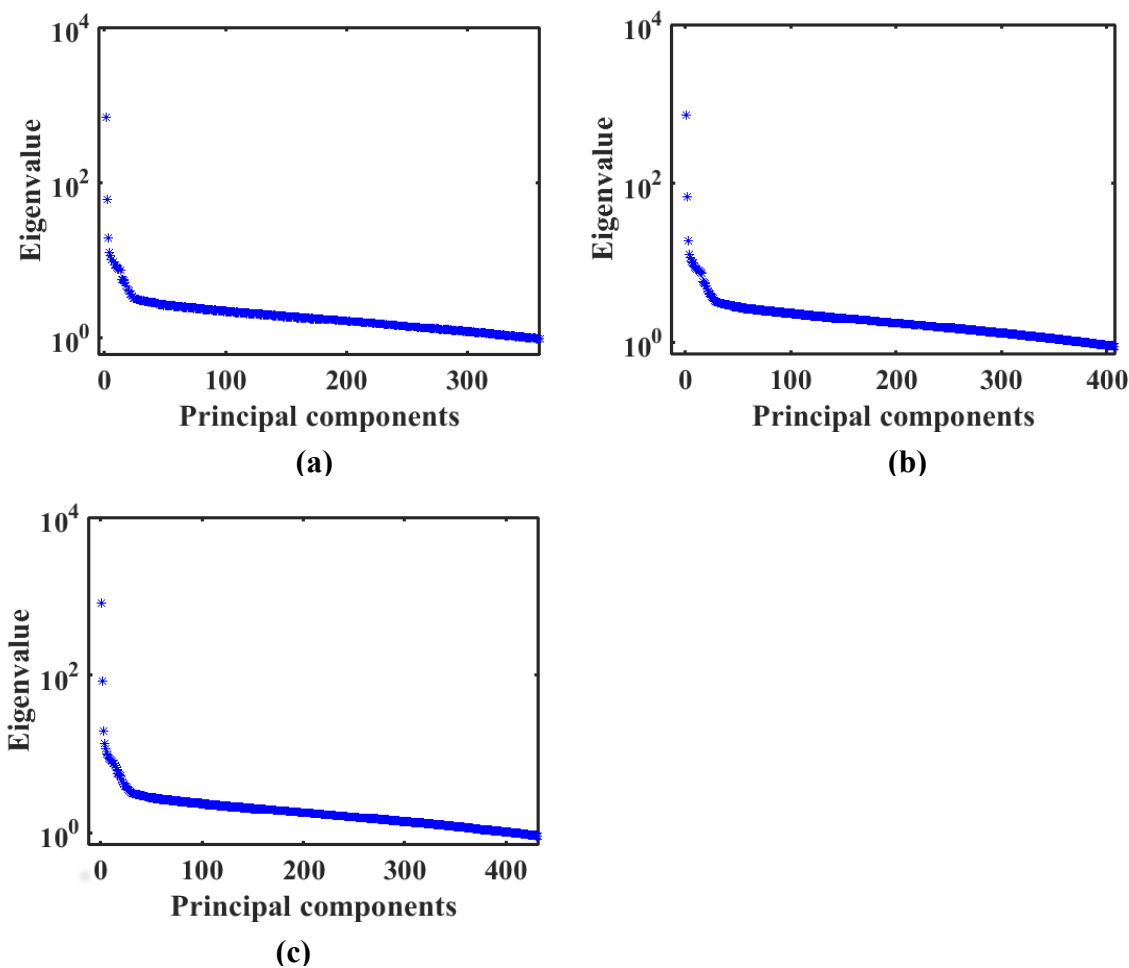


Figure 5.32 The log scree plots that explain overall scores discrimination in (a) grade₁, (b) grade₂ and (c) grade₃ spectral datasets of saliva samples from healthy volunteers (controls) and breast cancer patients.

Table 5.25 Categorization of PCs based on the cumulative percentage of total variation and the size of variances: Low-order PCs (<90% of the cumulative variance and >1.0 average eigenvalue), Intermediate-order PCs (between 90% and 95% of the cumulative variance), Higher-order PCs (>95% of the cumulative variance)

| Spectral dataset | Low-order PCs | Intermediate-order PCs | Higher-order PCs |
|--------------------|---------------|------------------------|------------------|
| Grade ₁ | 1-29 | 30-128 | 129-783 |
| Grade ₂ | 1-25 | 26-135 | 136-783 |
| Grade ₃ | 1-12 | 13-96 | 97-789 |

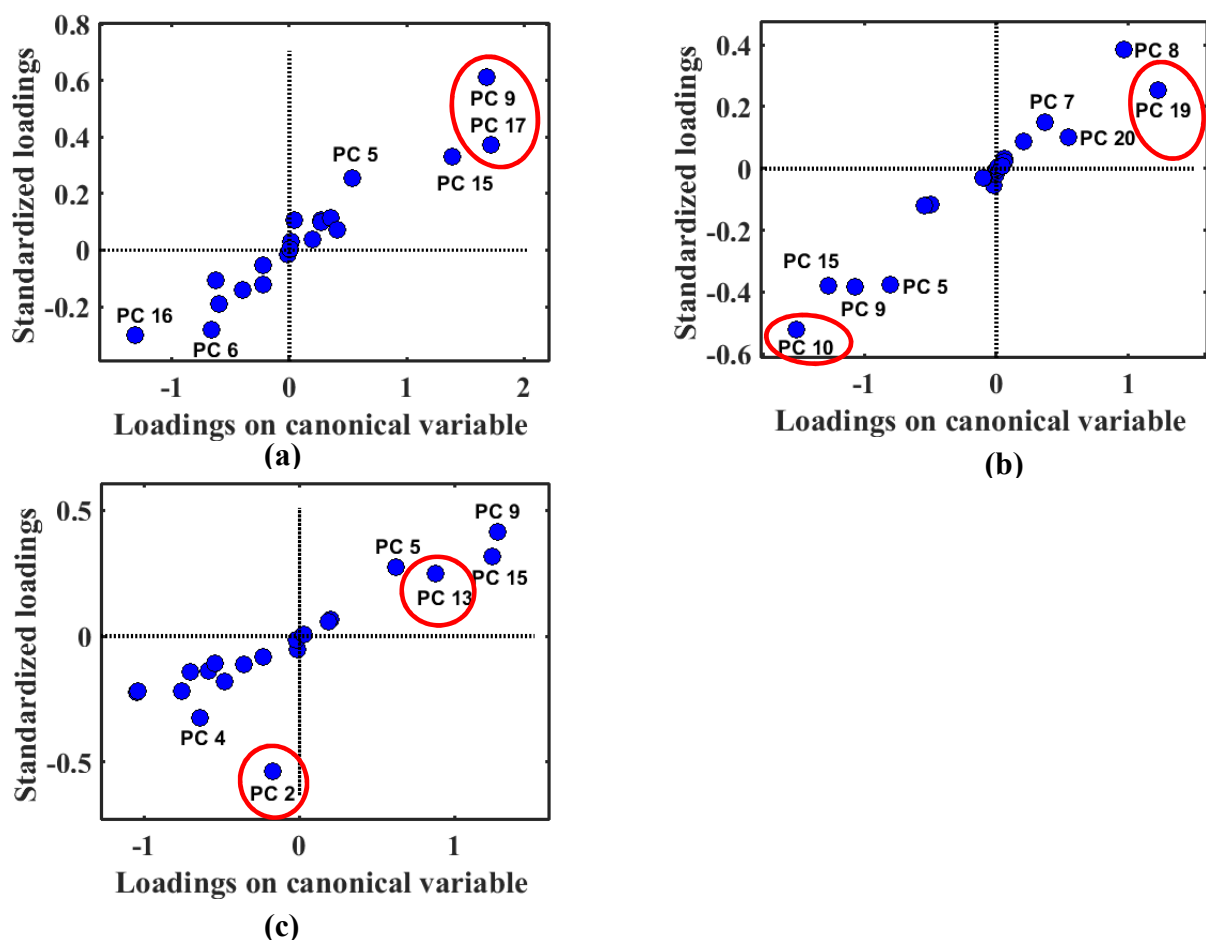


Figure 5.33 Canonical variable distribution showing the low- and intermediate- order principal components for (a) grade₁, (b) grade₂, and (c) grade₃ saliva spectral datasets, respectively. The PCs marked with a red circle were the most useful for scores discrimination.

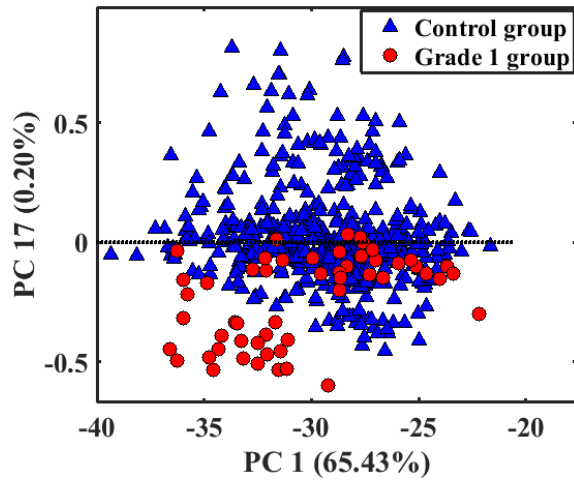
The loading vectors of PCs 9, 10 and 2 corresponded to the observed prominent peaks in Figure 5.31. The observed loading vectors were attributed to the major biochemical changes occurring during breast cancer progression. In view of the major objective being detection and quantification of subtle alterations occurring during cancer progression, the study analyzed the loading vectors of PCs 17, 19 and 13. Analysis showed that though PCs 17, 19 and 13 accounted for 0.20%, 0.18%, and 0.29% of the total variance in their respective input data, they had higher canonical loading parameters and statistical significances (Figure 5.33, Table 5.26). These PCs were chosen for further analysis in order to understand the associated subtle biochemical alterations occurring in breast cancer progression (Figure 5.34, Figure 5.35). Examination of loading functions (Figure 5.35 (a-c)) suggests that the subtle markers occurred at 643-647 cm^{-1} , 687-689 cm^{-1} , 816-818 cm^{-1} , 1022-1024 cm^{-1} , 1125-1128 cm^{-1} , 1145-1148 cm^{-1} , 1164-1166 cm^{-1} , 1427-1430 cm^{-1} , 1570-1572 cm^{-1} , 1609-1619 cm^{-1} , 1630-1657 cm^{-1} , and 1753-1756 cm^{-1} spectral regions. If we consider Figure 5.35 (a), it can be seen that the diseased samples had subtle loading vectors at 689, 1127, 1147, 1428, 1500, 1570 and 1643 cm^{-1} . In contrast, Figure 5.35 (b), and Figure 5.35(c) suggests that the diseased samples had strong loading vectors at (910, 1570, 1634 cm^{-1}) and (774, 966, 1657, 1710 cm^{-1}), respectively. In general, control samples exhibited heightened loading vectors at (643, 818, 1614, 1685 cm^{-1}), (816, 1276, 1479, 1609, 1683 cm^{-1}) and (1364, 1403, 1483 cm^{-1}), as observed in Figures 5.35(a), 5.35(b) and 5.35(c), respectively.

A statistical analysis on mean and standard deviations of common loading vectors were calculated on datasets of all stages of malignancy. The changes in ring breathing modes of DNA bases ($690 \pm 1.47 \text{ cm}^{-1}$), tyrosine proteins ($1159 \pm 4.24 \text{ cm}^{-1}$), amide II proteins ($1570 \pm 0.47 \text{ cm}^{-1}$), and amide I proteins ($1644 \pm 4.73 \text{ cm}^{-1}$) were observed to increase with breast cancer progression. In contrast, the changes in proline / tyrosine proteins (817 ± 0.40 , $1614 \pm 2.04 \text{ cm}^{-1}$) and lipids ($1754 \pm 0.23 \text{ cm}^{-1}$) were observed to decrease with cancer progression.

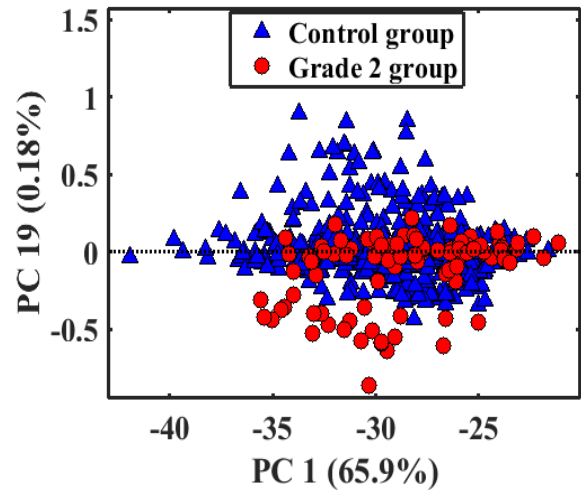
If we consider diseased samples (Figure 5.35 (a-c)), it can be seen that subtle changes in proteins ($1126 \pm 0.28 \text{ cm}^{-1}$) and CH_2 deformation ($1428 \pm 0.28 \text{ cm}^{-1}$) were detected in early stages of malignancy (grade 1 and grade 2), but could not be detected in late (grade 3) malignancy. Further, it is evident that there were subtle biochemical changes of C-C twisting mode of tyrosine ($645 \pm 1.15 \text{ cm}^{-1}$) in samples of control patients when compared to diseased patients. A similar comparison of loading vectors in controls versus grade 2 and grade 3 breast cancer patients suggested subtle biochemical changes due to amide III / collagen ($1281 \pm 3.17 \text{ cm}^{-1}$) and amide I ($1684 \pm 0.57 \text{ cm}^{-1}$) were prominent in samples from control patients.

Table 5.26 The statistical values (*t*-test (*p*-values) and effect sizes ((Cohen-*d*, Pearson's correlation coefficients (*r*)) showing relationship between the principal component scores of control and diseased saliva samples. For clarity, only the statistical values for statistically significant PCs (*p* < 0.05) are shown

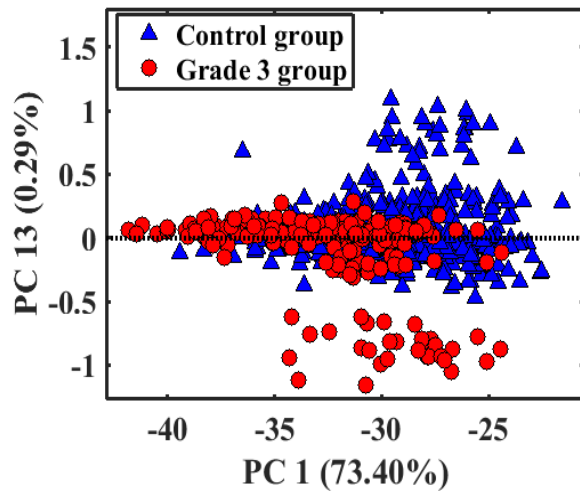
| PC | Grade 1 | | | Grade 2 | | | Grade 3 | | |
|----|-------------------------|-----------------|----------|-------------------------|-----------------|----------|-------------------------|-----------------|----------|
| | <i>p</i> -value | Cohen- <i>d</i> | <i>r</i> | <i>p</i> -value | Cohen- <i>d</i> | <i>r</i> | <i>p</i> -value | Cohen- <i>d</i> | <i>r</i> |
| 1 | 0.012 | -0.324 | -0.16 | 0.009 | 0.25 | 0.12 | 8.1 x 10 ⁻³¹ | -0.99 | -0.445 |
| 2 | 0.00223 | -0.4095 | -0.20 | <i>p</i> > 0.05 | | | 7.4 x 10 ⁻³¹ | 1.12 | 0.48 |
| 3 | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | |
| 4 | 0.0026 | 0.40 | 0.19 | <i>p</i> > 0.05 | | | 7.9 x 10 ⁻¹⁶ | 0.67 | 0.31 |
| 5 | 5.2 x 10 ⁻⁹ | -0.834 | -0.384 | 2.8 x 10 ⁻²⁰ | 1.051 | 0.465 | 1.9 x 10 ⁻¹² | -0.58 | -0.28 |
| 6 | 2.5 x 10 ⁻¹⁰ | 0.90 | 0.413 | 0.01 | -0.25 | -0.12 | 1.4 x 10 ⁻⁶ | 0.39 | 0.19 |
| 7 | 0.011 | -0.32 | -0.16 | 0.00018 | -0.39 | -0.194 | 0.013 | 0.12 | 0.09 |
| 8 | 0.014 | -0.314 | 0.15 | 4.7 x 10 ⁻²⁰ | -1.10 | -0.483 | <i>p</i> > 0.05 | | |
| 9 | 3.0 x 10 ⁻⁵¹ | -2.39 | -0.76 | 1.8 x 10 ⁻²⁰ | 1.05 | 0.467 | 5.2 x 10 ⁻²⁵ | -0.86 | -0.4 |
| 10 | 0.0013 | 0.43 | 0.21 | 2.0 x 10 ⁻³⁸ | 1.53 | 0.60 | 0.002 | 0.23 | 0.11 |
| 11 | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | |
| 12 | 0.0058 | -0.362 | -0.17 | <i>p</i> > 0.05 | | | 4.0 x 10 ⁻⁸ | 0.44 | 0.21 |
| 13 | 1.3 x 10 ⁻⁵ | 0.608 | 0.29 | <i>p</i> > 0.05 | | | 2.5 x 10 ⁻¹⁰ | -0.52 | -0.25 |
| 14 | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | | <i>p</i> > 0.05 | | |
| 15 | 2.8 x 10 ⁻¹⁴ | -1.10 | 0.48 | 2.5 x 10 ⁻²⁰ | 1.05 | 0.465 | 4.6 x 10 ⁻¹⁵ | -0.65 | -0.31 |
| 16 | 2.2 x 10 ⁻¹² | 1.01 | 0.45 | 0.00512 | 0.28 | 0.14 | 0.00056 | 0.27 | 0.13 |
| 17 | 1.3 x 10 ⁻¹⁷ | -1.254 | 0.53 | <i>p</i> > 0.05 | | | 2.4 x 10 ⁻⁸ | 0.45 | 0.22 |
| 18 | <i>p</i> > 0.05 | | | 0.00288 | 0.30 | 0.15 | 2.0 x 10 ⁻⁸ | 0.459 | 0.22 |
| 19 | <i>p</i> > 0.05 | | | 6.9 x 10 ⁻¹⁰ | -0.68 | -0.32 | 0.00014 | 0.30 | 0.14 |
| 20 | 0.007 | 1.411 | 0.57 | 0.00835 | -0.984 | -0.44 | 0.00379 | 0.54 | 0.26 |



(a)



(b)



(c)

Figure 5.34 Scatter plots showing distribution of low- order PC (PC 1) scores versus (a) PC 17 scores, (b) PC 19 scores, and (c) PC 13 scores.

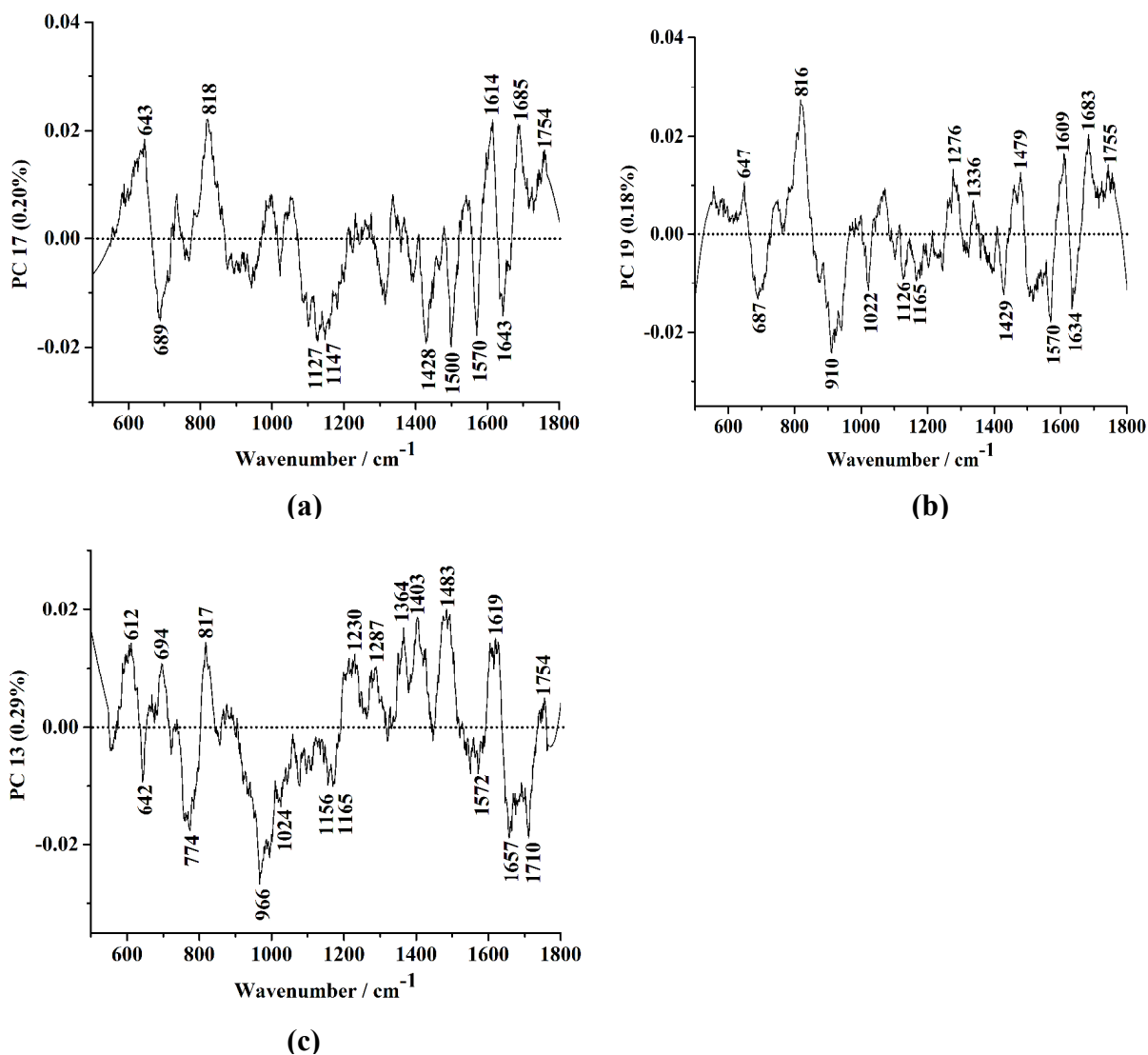


Figure 5.35 Loading functions explaining scores discrimination of control and diseased saliva spectra. It is observed that subtle band alterations mainly featured at 643-647 cm^{-1} , 687-689 cm^{-1} , 816-818 cm^{-1} , 1022-1024 cm^{-1} , 1125-1128 cm^{-1} , 1145-1148 cm^{-1} , 1164-1166 cm^{-1} , 1427-1430 cm^{-1} , 1570-1572 cm^{-1} , 1609-1619 cm^{-1} , 11630-1657 cm^{-1} , and 1753-1756 cm^{-1} spectral regions.

5.2.2.3 Quantitative analysis of trace biomarkers in saliva spectra using partial least-squares regression

The calibration set design for biochemical components formulation in simulate saliva samples has been previously provided (Table 4.4). Spectral matrices collected from calibration samples at spectral regions 643-647 cm^{-1} (tyrosine, phenylalanine), 687-689 cm^{-1} (DNA), 816-818 cm^{-1} (proline, hydroxyproline, tyrosine, collagen), 1022-1024 cm^{-1} (glycogen), 1125-1128 cm^{-1} (lipids, proteins), 1145-1148 cm^{-1} (glycogen, carotenoids), 11164-11166 cm^{-1} (tyrosine), 1427-1430 cm^{-1} (deoxyribose, lipids), 1570-1572 cm^{-1} (DNA), 1609-1619 cm^{-1} (cytosine, tyrosine, phenylalanine, tryptophan), 11630-1657 cm^{-1} (amide I), and 1753-1756 cm^{-1} (lipids) were used for model calibration. Based on the root-mean-square error of cross-validation (R^2_{val}) (Høy *et al.*, 2012), we observed the predicted versus measured regression plots suggested that the model worked well for RNA and glutamate components followed by glycine, glycerol, albumen, glycogen and triolein components (Figure 5.36). The calculated biochemical concentrations in a standard saliva simulate using the PLS regression model showed the concentration levels were in agreement within $\pm 10\%$ deviation (Table 5.27). Regarding the limits of detection and the limit of quantification of the PLS models (Table 5.28) it was verified that the PLS model could detect amounts within the expected ranges in human body (Table 1) and $R^2 > 0.89$. As shown by other findings (Saeys *et al.*, 2005), a calibration model with R^2_{val} value greater than 0.91 is considered to be an excellent calibration, while an $R^2 > 0.82$ results in good prediction (Suhandy *et al.*, 2012). As the concentration of biochemical components in the proposed PLS model ranged from 1-500 ppm, the model was effective to detect and quantify the trace nucleic acids, proteins, lipids and saccharides in the human body. For analysis, trace biomarkers in saliva were quantified in the 'fingerprint' (500-1800 cm^{-1}) region (Table 5.29 (a)) and in the selected 643-647, 687-689, 816-818, 1022-1024, 1125-1128, 1145-1148, 11164-11166, 1427-1430, 1570-1572, 1609-1619, 11630-1657, 1753-1756 cm^{-1} regions (Table 5.29 (b)). For plotting, the determined concentration levels (in ppm) were normalized to their mean value and the levels correlated to cancer presence and severity.

Table 5.27 Comparison of biochemical components concentrations in a saliva simulate reference solution and the results obtained from PLS regression

| Biochemical Components | Concentration (mg / ml) | Measured value (\pm SD) | Deviation (%) |
|------------------------|-------------------------|----------------------------|---------------|
| Albumen | 0.4 | 0.42 \pm 0.02 | 5 |
| Glycogen | 0.1 | 0.11 \pm 0.025 | 10 |
| Glutamate | 0.001 | 0.000971 \pm 0.00012 | 2.9 |
| Glycerol | 0.01 | 0.0104 \pm 0.0024 | 4 |
| RNA | 0.002 | 0.0021 \pm 0.00011 | 5 |
| Triolein | 0.3 | 0.303 \pm 0.0036 | 1 |

Table 5.28 Detection limits (mg/ml) of biochemical components for Raman analysis of simulate saliva

| Biochemical component | Detection limits | | | |
|-----------------------|------------------------|-----------------------|-------------------------|-------|
| | LOD | LOQ | (<i>RMSEP</i>) | R^2 |
| Albumen | 0.0089 | 0.027 | 0.00172 | 0.996 |
| Glycogen | 0.0234 | 0.073 | 0.00211 | 0.984 |
| Glutamate | 1.669*10 ⁻⁸ | 5.08*10 ⁻⁸ | 1.415*10 ⁻¹⁰ | 1 |
| Glycerol | 0.0022 | 0.0067 | 0.00431 | 0.998 |
| Glycine | 0.00168 | 0.00532 | 0.00172 | 0.997 |
| RNA | 0.00144 | 0.0043 | 1.178*10 ⁻¹¹ | 1 |
| Triolein | 0.0481 | 0.145 | 0.0077 | 0.898 |

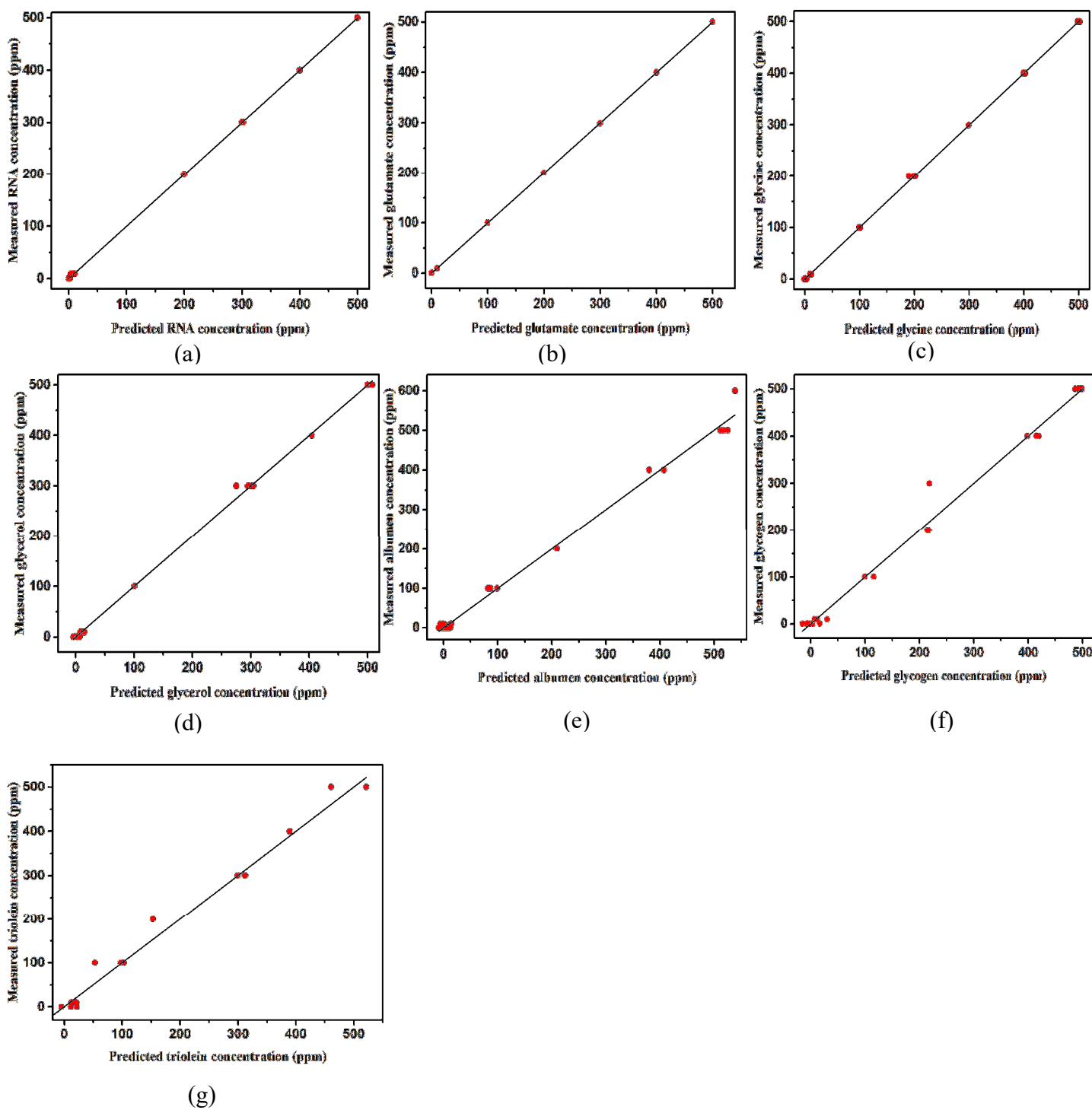


Figure 5.36 Regression plots for partial least squares measured versus predicted biochemical concentrations of the pure biochemical compounds (a) RNA, (b) glutamate, (c) glycine, (d) glycerol, (e) albumen, (f) glycogen, and (g) triolein.

Table 5.29 Estimated amounts of biochemical components in saliva of control and breast cancer patients in fingerprint region (500-1800 cm^{-1}) and the selected subtle (643-647, 687-689, 816- 818, 1022-1024, 1125-1128, 1145-1148, 11164-1166, 1427-1430, 1570-1572, 1609-1619, 11630-1657, 1753-1756 cm^{-1}) spectral regions

(a)

| 500-1800 cm^{-1} region | | Biochemical components (ppm) | | | | | |
|----------------------------------|---------|------------------------------|-----------|----------|---------|------|----------|
| Disease status | Albumen | Glycogen | Glutamate | Glycerol | Glycine | RNA | Triolein |
| Controls | 50 | 33.9 | 3.61 | 247.7 | 2.05 | 4.82 | 43.8 |
| Diseased | 526.4 | 8.3 | 14.3 | 359.2 | 6.30 | 7.31 | 91.5 |

(b)

| Based on subtle band regions | | Biochemical components (ppm) | | | | | |
|------------------------------|---------|------------------------------|-----------|----------|---------|------|----------|
| Disease status | Albumen | Glycogen | Glutamate | Glycerol | Glycine | RNA | Triolein |
| Controls | 27.7 | 33.9 | 3.66 | 145.2 | 2.17 | 7.35 | 43.9 |
| Grade 1 | 62.5 | 13.1 | 9.1 | 147.6 | 19.6 | 7.60 | 45.7 |
| Grade 2 | 78.9 | 12.7 | 5.78 | 150.7 | 18.2 | 8.29 | 59.5 |
| Grade 3 | 126.3 | 11.5 | 4.90 | 359.6 | 20.6 | 9.16 | 61.8 |

It can be seen (Table 5.29(a)) that the relative amounts of glycogen biochemical components were greater in control patients when compared to diseased patients, meaning that the total amounts of saccharides were greater in control patients, which is in agreement with a previous biochemical-cytological study (Emekli-Altufran *et al.*, 2008). In contrast, the relative amounts of albumen, glutamate, glycine, RNA, glycerol and triolein biochemical components were greater in diseased patients when compared to control patients which suggest that proteins, amino acids, nucleic acids and lipids were greater in diseased patients. If we consider Table 5.29 (b), it is seen that levels of glycogen decreased with progression of malignancy whereas levels of albumen, RNA, glycerol and triolein increased with malignancy. The decrement of glycogen content with malignancy can be associated with enhanced glucose uptake by cells during onset of tumor development for conversion to lactate molecules necessary for energy production during cell proliferation (Klement *et al.*, 2013). Other studies (Gonchukov *et al.*, 2012) have shown that elevated nucleic acids in saliva can be used as a biomarker for cancer progression. Moreover, spectral intensities of DNA and RNA related bands have been observed to increase with other malignancies (Stone *et al.*, 2007; Taleb *et al.*, 2006; Crow *et al.*, 2003) a factor attributed to

abundance of DNA content in malignant samples (Taleb *et al.*, 2006). The heightened adipocyte levels of membranous lipids in saliva of diseased patients can be associated with production of lipids via *de novo* lipogenesis (Long *et al.*, 2018).

The ultimate diagnostic classification of each Raman spectrum was determined by PLS-DA with the $k = 10$ fold cross-validation method. Figure 5.37 (a-e) shows the scatter plots of the linear discriminant analysis, demonstrating the clustering of biochemical alterations of saliva from the normal patients (controls) and malignant breast tumor categories / patients using the PLS-DA diagnostic algorithm. We find a clear separation between the grade 1 and grade 2 scores, and between grade 2 and grade 3 scores groups was achieved, primarily by the first three discriminant functions. Similarly, a separation of controls and malignant scores could also be observed, although some larger overlap existed, which presumably reflected that there were similar alterations in the saliva components in controls' saliva and diseased patients' saliva.

The stage-wise comparison of breast cancer (Figure 5.38 – Figure 5.40) suggests that there were mixed biochemical alterations that could be associated with stages of cancer progression. For instance (Figure 5.38 (a)), changes in RNA / amino acids (814 cm^{-1}), glycogen (1024 cm^{-1}), CH_2 deformation (1427 cm^{-1}) were dominant in control samples when compared to diseased samples that had heightened alterations of C=N adenine (1568 cm^{-1}), cytosine (1609 cm^{-1}), and lipids (1754 cm^{-1}). Similarly (Figure 5.38 (b)), biochemical changes due to C–C twisting mode of tyrosine (646 cm^{-1}), DNA bases (686 cm^{-1}), collagen, proteins (818 cm^{-1}) were detected in control samples in comparison to diseased samples that were dominated by changes in lipids (1754 cm^{-1}). In Figure 5.39 (a), it can be seen that biochemical changes due to proteins / saccharides (1125 cm^{-1}), amino acids (1162 cm^{-1}), and lipids / tyrosine (1168 cm^{-1}) were prominent in control samples whereas changes in tyrosine (646 cm^{-1}), tryptophan (1620 cm^{-1}), collagen / amide I ($1635, 1640\text{ cm}^{-1}$) and lipids (1752 cm^{-1}) were dominant in late (stage 3) malignancy.

As seen in Figure 5.39 (b), the intense negative loadings shows that subtle changes in urea / triglycerides (1403 cm^{-1}), and adenine / amide II (1568 cm^{-1}) were detected in grade 1 breast cancer, whereas changes in tyrosine (643 cm^{-1}), proline / tyrosine (814 cm^{-1}), proteins (819 cm^{-1}) and glycogen (1024 cm^{-1}) were detected in grade 2 breast cancer. Comparison of breast cancer grade -2 and breast cancer grade -3 scores was performed by examining the third latent variable (LV 3) as shown in Figure 5.40. It is seen the Raman spectral features associated with breast cancer stage-2 included 814 cm^{-1} (proline / tyrosine), 819 cm^{-1} (proteins), 1167 cm^{-1} (lipids, tyrosine) and 1608 cm^{-1} (cytosine), whereas subtle biochemical changes due to adenine (686 cm^{-1}), adenine

(1144 cm^{-1}), cytosine / adenine (1610 cm^{-1}), amide I (1629 cm^{-1}), amide I (1635 cm^{-1}), amide I (1640 cm^{-1}), amide I (1646 cm^{-1}) and amide I (1655 cm^{-1}) were detected in late malignancy (grade -3 cancer).

To better understand changes in malignancy, the peak ratios i.e. (I_C/I_N) were determined, where the normalized spectra peak intensities of diseased samples (I_C) were divided by the normalized spectra peak intensities of control samples (I_N). The peak ratios (I_C/I_N) due to changes in tyrosine proteins (642-648 cm^{-1} , 815-817 cm^{-1} , 1165 cm^{-1}), DNA (689 cm^{-1} , 815-817 cm^{-1} , 1427 cm^{-1} , 1620 cm^{-1}), amide I (1631-16320 cm^{-1} , 1652 cm^{-1}), and lipids (1752 cm^{-1}) were found to increase with breast malignancy, suggesting the potential of the saliva proteins and nucleic acids in breast cancer detection and screening.

The developed PLS-DA algorithm achieved diagnostic sensitivities of 93%, 91%, and 91%; specificities of 96%, 93%, and 91%; and accuracies of 96%, 92%, and 91%, respectively, when differentiating normal saliva samples from grade 1 breast cancer in saliva samples, grade 2 breast cancer in saliva samples, and grade 3 breast cancer in saliva samples (Table 5.30). These diagnostic performances (> 90%) demonstrate that the PLS-DA-based saliva Raman spectral classification method is powerful for the differentiation of different stages of breast cancer.

As a potential diagnostic media for disease detection, the biochemical composition of human saliva may be closely related to metabolic abnormalities when disease afflicts the body, which makes it possible to detect many diseases via the Raman spectral features of saliva. However, there are very few reports using regular Raman spectroscopy to study human saliva samples for cancer detection (Calado *et al.*, 2019; Scott *et al.*, 2010; Gonchukov *et al.*, 2012; Li *et al.*, 2012), due to its inherently small scattering cross-section and the strong background fluorescence interference (Feng *et al.*, 2015). These limitations of regular Raman most likely make the technique not sensitive enough for detecting the subtle biochemical changes in human saliva samples for medical diagnosis (Feng *et al.*, 2015).

With regard to breast cancer, utility of saliva proteins for the noninvasive differentiation of benign and malignant breast tumors has been recently reported (Feng *et al.*, 2015; Wu *et al.*, 2015). In the present study, we have observed that relative amounts of proteins, amino acids, nucleic acids and lipids were greater in the saliva of diseased patients whereas the the total amounts of saccharides were greater in saliva of control patients (Table 5.29 (a), (b)). The abundance of total amount of proteins and the abundance of tyrosine proteins (642-648 cm^{-1} , 815-817 cm^{-1} , 1165 cm^{-1}) and amide (1631-16320 cm^{-1} , 1652 cm^{-1}) components agrees with related works

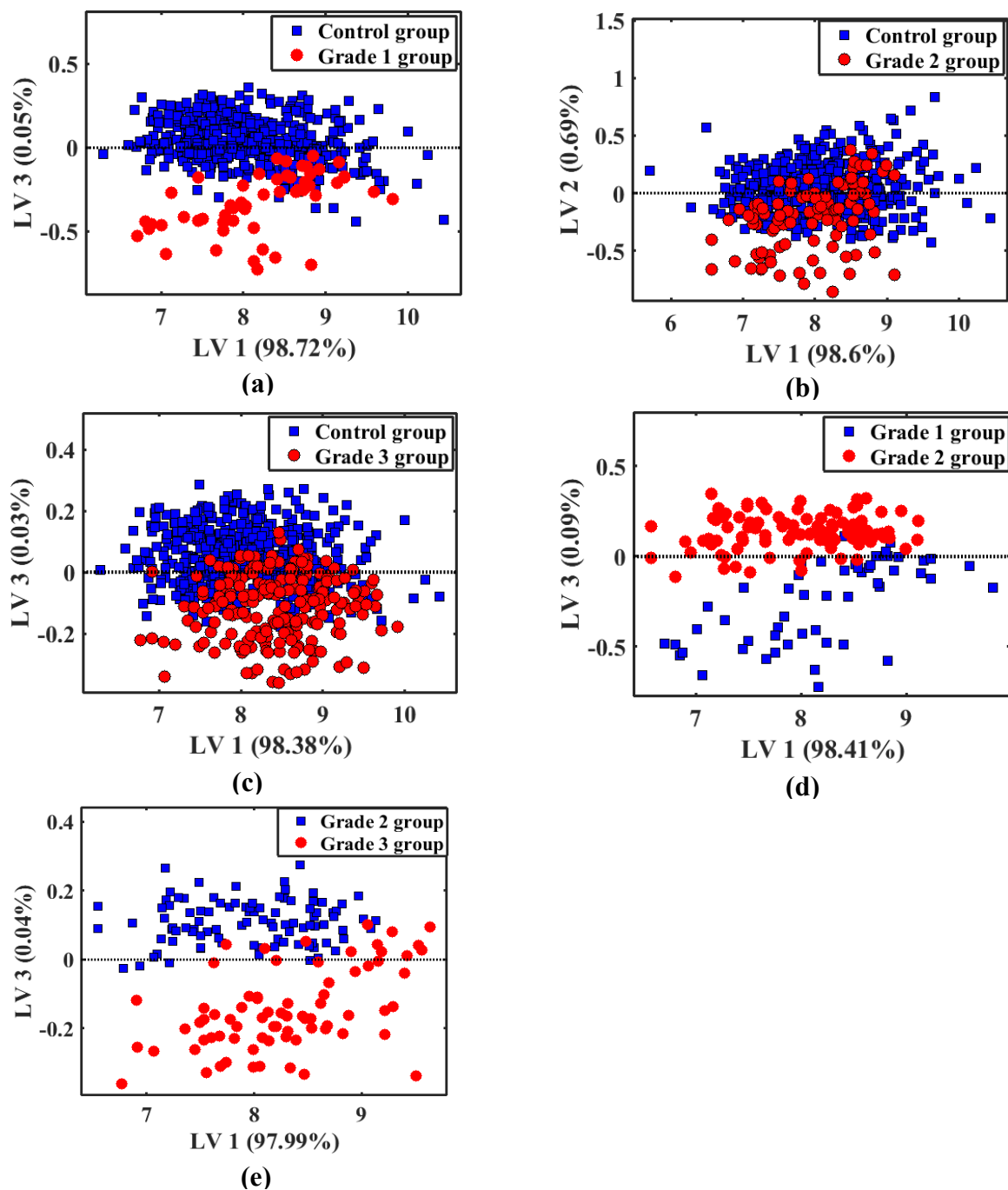
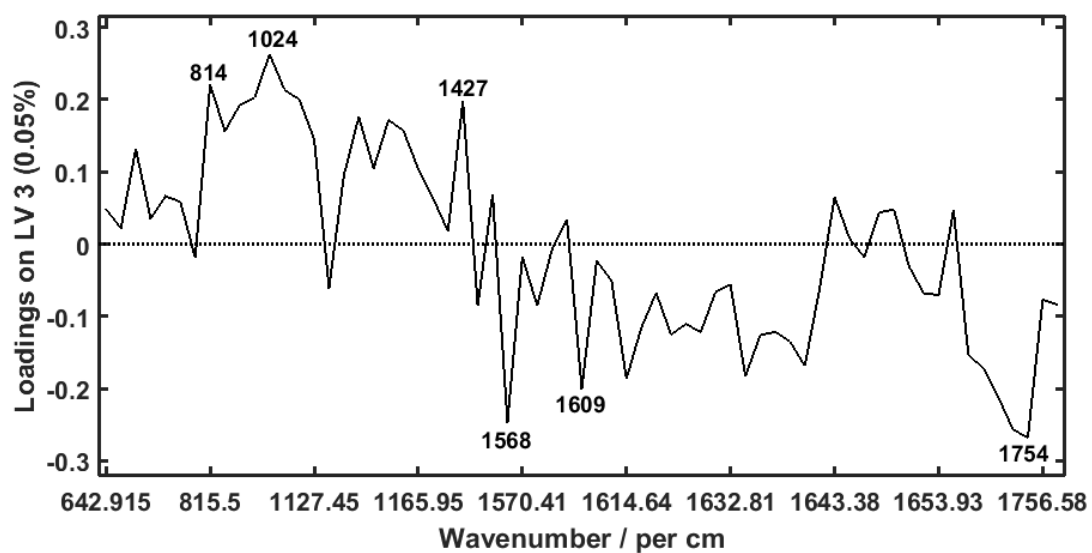
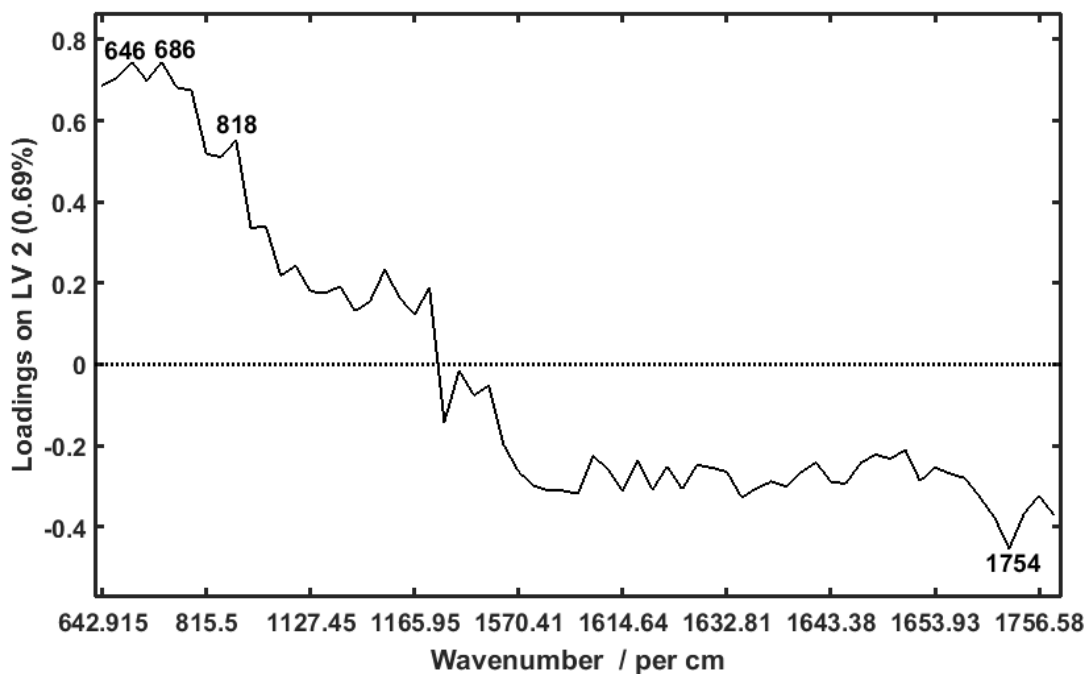


Figure 5.37 PLS-DA scatterplots showing differentiation of (a) controls from grade 1, (b) controls from grade 2, (c) controls from grade 3, (d) grade 1 from grade 2, and (e) grade 2 from grade 3 breast cancers; based on the spectra regions: 643-647, 687-689, 816-818, 1022-1024, 1125-1128, 1145-1148, 11164-1166, 1427-1430, 1570-1572, 1609-1619, 11630-1657, and 1753-1756 cm^{-1} .

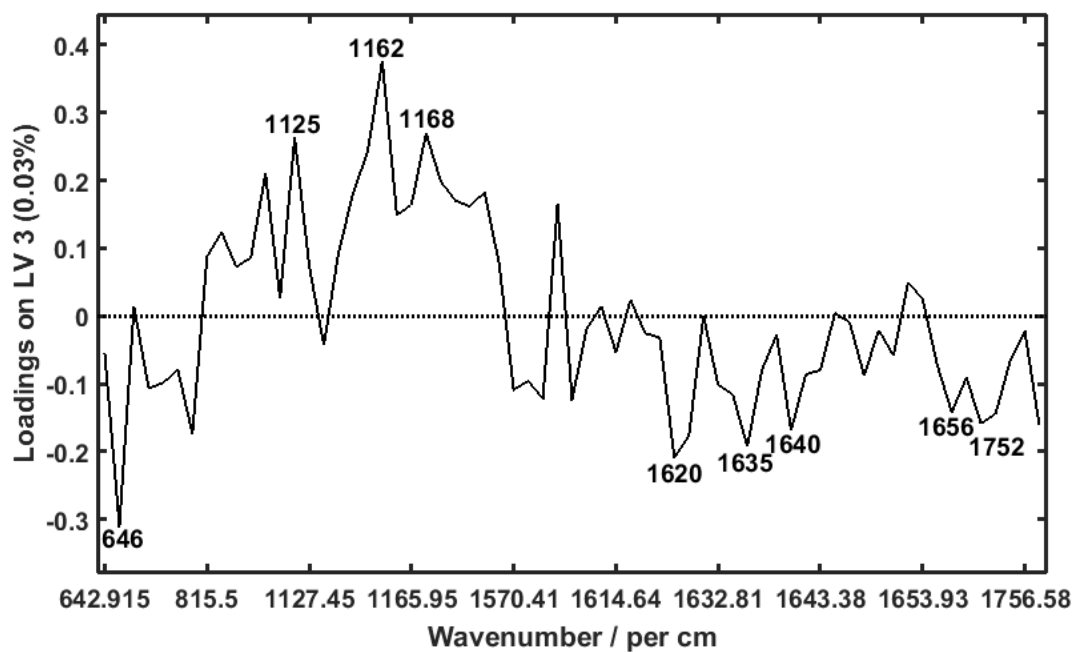


(a)

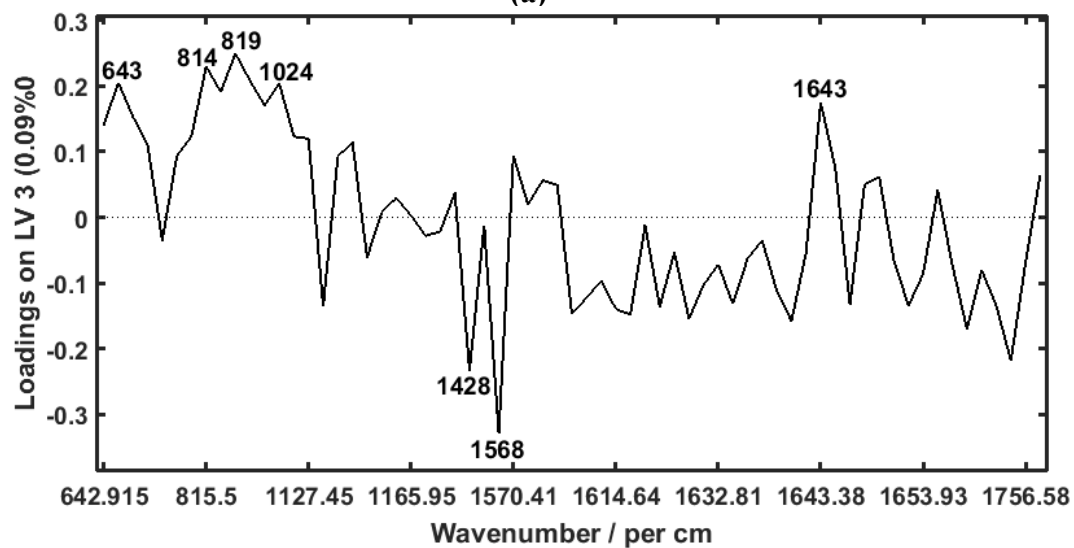


(b)

Figure 5.38 Loading functions explaining differentiation of (a) controls from grade 1 breast cancer, and (b) controls from grade 2 breast cancer.



(a)



(b)

Figure 5.39 Loading functions explaining differentiation of (a) controls from grade 3 breast cancer scores, and (b) grade 1 breast cancer from grade 2 breast cancer.

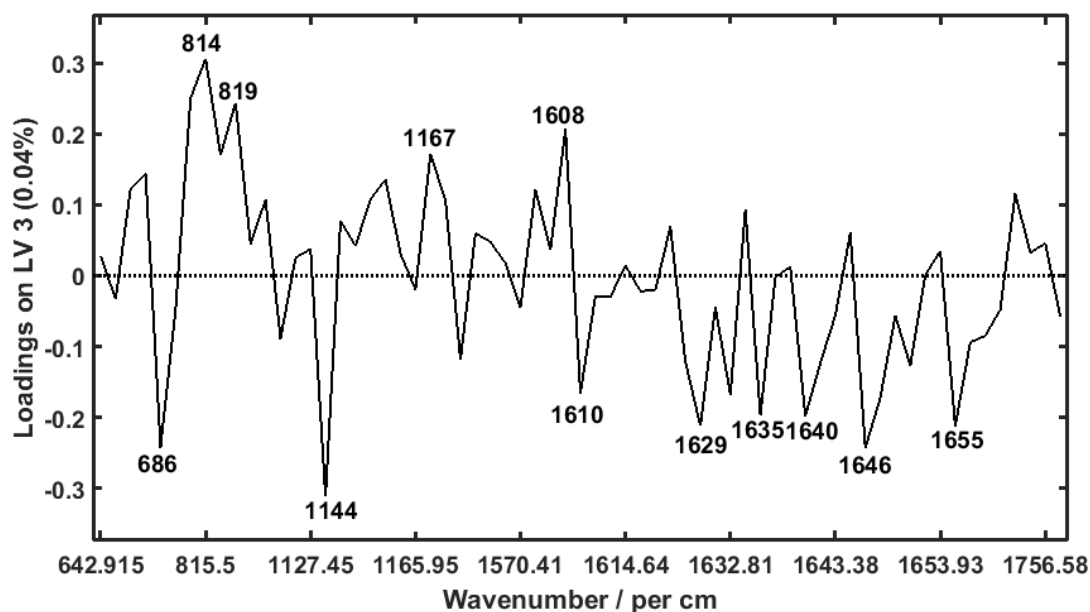


Figure 5.40 Loading functions explaining differentiation of grade 2 breast cancer from grade 3 breast cancer.

Table 5.30 Diagnostic results of PLS-DA on the Raman spectra of saliva from healthy volunteers (controls) and breast cancer patients

| | | Cases | | | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|----------------|---------------|---------------|----------|--------------|-------|--------------|-----------------|-----------------|
| Disease status | Diagnosis | Breast cancer | Controls | Not assigned | | | | |
| Grade-1 | Breast cancer | 50 | 2 | 2 | 54 | 96% | 93% | 96% |
| | Controls | 11 | 483 | 11 | 505 | | | |
| Grade-2 | Breast cancer | 88 | 6 | 3 | 97 | 92% | 91% | 93% |
| | Controls | 18 | 476 | 20 | 514 | | | |
| Grade-3 | Breast cancer | 188 | 14 | 6 | 208 | 91% | 91% | 91% |
| | Controls | 25 | 457 | 17 | 499 | | | |

(Feng *et al.*, 2015; Wu *et al.*, 2015) where the significant Raman peaks corresponding to amide III and amide I presented higher Raman signals in malignant breast cancer, which can be attributed to vibrational modes of the amino acid bonds of the secondary structure of proteins (Movasaghi *et al.*, 2007).

Elsewhere, Ferreira *et al.*, (2020) observed changes in protein components by employing attenuated total reflection-Fourier transform infrared (ATR-FTIR) spectroscopy on saliva samples to discriminate breast cancer patients from benign patients and healthy subjects. The absorbance levels were observed to be significantly higher in saliva of breast cancer patients compared with benign patients at wavenumber 1041 cm^{-1} (collagen proteins) and the ROC curve analysis of this peak showed a reasonable accuracy to discriminate breast cancer from benign and control patients. Moreover, the $1433\text{--}1302.9\text{ cm}^{-1}$ wavenumber region (assigned to $\text{CH}_2 / \text{CH}_3$ wagging, twisting and bending modes of collagen and lipids) was elevated in saliva of breast cancer patients when compared to control and benign patients. It is thought that due to the desmoplastic reaction, the deposition of abundant collagen will occur as a stromal response to breast carcinoma, which may be reflected in the saliva protein spectra (Feng *et al.*, 2015). Hence, the protein signals from saliva samples as biomarkers observed between healthy and diseased subjects indicate that saliva protein Raman spectra can be employed to elucidate biomolecular and inherent changes of breast tumor subjects (Feng *et al.*, 2015). Indeed, the exploitation of protein-specific Raman signals from saliva samples has also been put forward as a potential method of studying biomarkers for nasopharyngeal cancer screen (Feng *et al.*, 2014; Qiu *et al.*, 2016).

In the present study, the peaks centered at 689 cm^{-1} , $815\text{--}817\text{ cm}^{-1}$, 1427 cm^{-1} and 1620 cm^{-1} which are the ring breathing modes of adenine (Movasaghi *et al.*, 2007), demonstrate a higher intensity in the diseased patients, indicating that the amount of RNA in the saliva of diseased patients was increased. This can be attributed to circulating nucleic acid contents mainly resulting from apoptosis of carcinoma cells and necrosis in the tumor microenvironment (Qiu *et al.*, 2016). With regard to lipids, tumor cells are known to demonstrate extremely high endogenous fatty acid synthesis, regardless of the level of circulating fatty acids (Qiu *et al.*, 2016). Circulating fatty acids may directly promote tumor cell growth and metastasis (Bauer *et al.*, 2005) suggesting that breast cancer patient's saliva may be associated with an increased levels of fatty acids.

5.2.2.4 Multivariate exploratory analysis of Independent Component Analysis (ICA), Multidimensional Scaling (MDS), Partial least Square Discriminant Analysis (PLS-DA) and kernel density estimators for breast cancer diagnostics in saliva

ICA by Maximum Likelihood (ML) fast fixed-point estimation algorithm on spectral matrices of 643-647, 687-689, 816-818, 1022-1024, 1125-1128, 1145-1148, 11164-1166, 1427-1430, 1570-1572, 1609-1619, 11630-1657, 1753-1756 cm^{-1} regions shows the data could be majorly accounted for by 20 eigenvalues (Figure 5.41). Therefore, 20 eigenvalues, accounting for more than 90% variance were selected for further analysis (Table 5.31). Compared to number of eigenvalue determined by ICA analysis on blood datasets (Figure 5.21, Table 5.15), it can be concluded that saliva datasets were transformed into many directions of new feature spaces (and therefore magnitudes), potentially suggesting that spectral biochemical components of saliva were more complex by nature in comparison to components in blood samples.

PLS-DA was performed on corresponding coefficients of the combinations decomposed by ICA. In total, five latent variables (LVs) i.e., LV 1 to LV 5 were revealed to be the most diagnostically significant ($P < 0.05$) for detecting breast cancer patients. Based on various combinations of significant LVs, the LV 1 versus LV 2 scatter plots were generated to compare breast cancer patients and healthy volunteers (Figure 5.42). The breast cancer patient scores (circles) and healthy volunteer scores (triangles) were distributed in 2 separate directions, and the distribution of the healthy volunteer scores was more compressed than that of breast cancer patients.

Table 5.31 Selected dimensions (eigenvalues) and explained total variances for ICA by Maximum Likelihood (ML) fast fixed-point estimation on Raman spectra of saliva samples from healthy volunteers (control) and breast cancer patients

| Datasets | Dimensions | Sum of eigenvalues retained (%) |
|----------|------------|---------------------------------|
| Grade 1 | 20 | 90.22% |
| Grade 2 | 20 | 90.78% |
| Grade 3 | 20 | 92.47% |

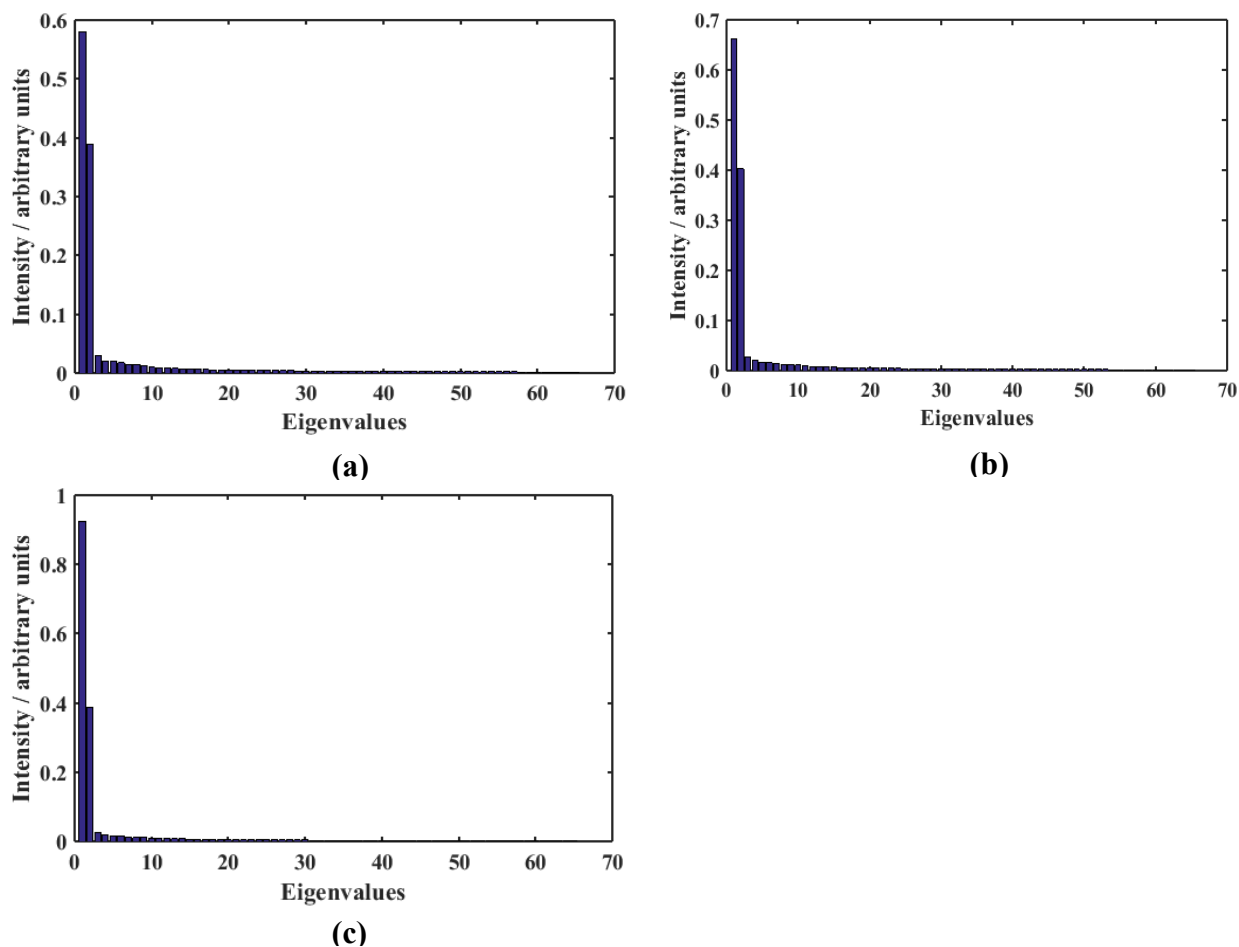


Figure 5.41 The eigenvalues for Raman spectra of saliva samples from healthy volunteers and (a) grade 1, (b) grade 2, and (c) grade 3 breast cancer patients.

For grade 1 datasets, the loadings on latent variables (Figure 5.42 (b, d, f)) shows the the fifth (IC5) and twelfth (IC12) independent components dominantly explained clustering of diseased and control scores, respectively. The fourteenth (IC14) and eighth (IC8) independent components explained the spectral regions that influenced clustering of diseased and control scores in grade 2 dataset, respectively. Four independent components were instrumental for clustering scores in stage 3 dataset. The tenth (IC10) and fourteenth (IC14) independent components greatly influenced clustering of control scores, whereas the first (IC1) and nineteenth (IC19) greatly influenced clustering of diseased scores. The respective positive and negative peaks are provided in Figure 5.43 to Figure 5.44. To better comprehend the molecular basis for the observed independent components' positive and negative peaks (Figures 5.43 - 5.44), the tentative

assignments of the Raman bands were done according to the known literature (Movasaghi *et al.*, 2007; Gelder *et al.*, 2007).

It can be seen (Figure 5.43 a, b and Figure 5.43 c, d) that biochemical changes due to C-C twisting mode of phenylalanine proteins (647), collagen (1165) and amides (1636 / 1641) were detected in saliva of healthy volunteers and breast cancer patients. However, analysis showed that subtle but significant changes in nucleic acids (689 cm^{-1}), proline / tyrosine proteins (816 cm^{-1}) nucleic acids / proteins (1573 cm^{-1}) were dominant in saliva of healthy volunteers scores whereas spectral markers associated with phenylalanine proteins (643 cm^{-1}), glycogen (1023 cm^{-1}), nucleic acids (1425 cm^{-1} , 1607 cm^{-1}), amide I (1630 cm^{-1} , 1647 cm^{-1}) influenced clustering of saliva spectra of breast cancer patients. In comparison to the breast cancer patients suffering from grade 2 breast cancer, biochemical changes due to lipids / proteins (1123 cm^{-1} , 1165 cm^{-1} , 1607 cm^{-1}), glycogen (1144 cm^{-1}), nucleic acids (1427 cm^{-1} , 1607 cm^{-1}) and 1638 cm^{-1} (amide I) were dominant in saliva collected from healthy volunteers. In contrast, spectral markers assigned to proline / tyrosine proteins (816 cm^{-1}), nucleic acids (1569 cm^{-1}) and amide I (1649 cm^{-1}) were dominantly detected in saliva of breast cancer patients.

For discrimination of saliva collected from healthy volunteers and patients suffering from late malignancy (Figure 5.44 (a-d)), it can be observed there were various influential prominent biochemical changes due to C-C twisting mode of phenylalanine proteins (645 / 647 cm^{-1}), proline / tyrosine (815 / 818 cm^{-1}), glycogen (1023 cm^{-1}), lipids / proteins (1123 / 1125 cm^{-1}), nucleic acids (1425 – 1428 cm^{-1} , 1567 / 1571 cm^{-1}), phenylalanine / tyrosine (1617 / 1620 cm^{-1}), and amide I (1645 / 1649 cm^{-1}). Further, the spectral marker assigned to amide I (1632 cm^{-1}) was dominant for clustering of saliva spectra from healthy volunteers, whereas the spectral markers assigned to carotenoids (1144 cm^{-1} , 1148 cm^{-1}), lipids / carotenoids (1165/67 cm^{-1}), nucleic acids (1608 cm^{-1}), and amide I (1640 cm^{-1}) uniquely determined clustering of diseased scores.

A diagnostic sensitivity of 89%, 95% and 92%, with a specificity of 95%, 95% and 92%, were achieved for the breast cancer patients and the healthy volunteers, respectively (Table 5.32), when the separation lines, which classified breast cancer patients from healthy volunteers, were set in Figure 5.42 (a, c, e). Nevertheless, it can be seen (Table 5.32) that various scores (grade 1 = 13; grade 2 = 18; grade 3 = 27) could not be assigned to either of the classes i.e., controls or diseased class, and could not therefore be used for saliva spectra discrimination, which potentially

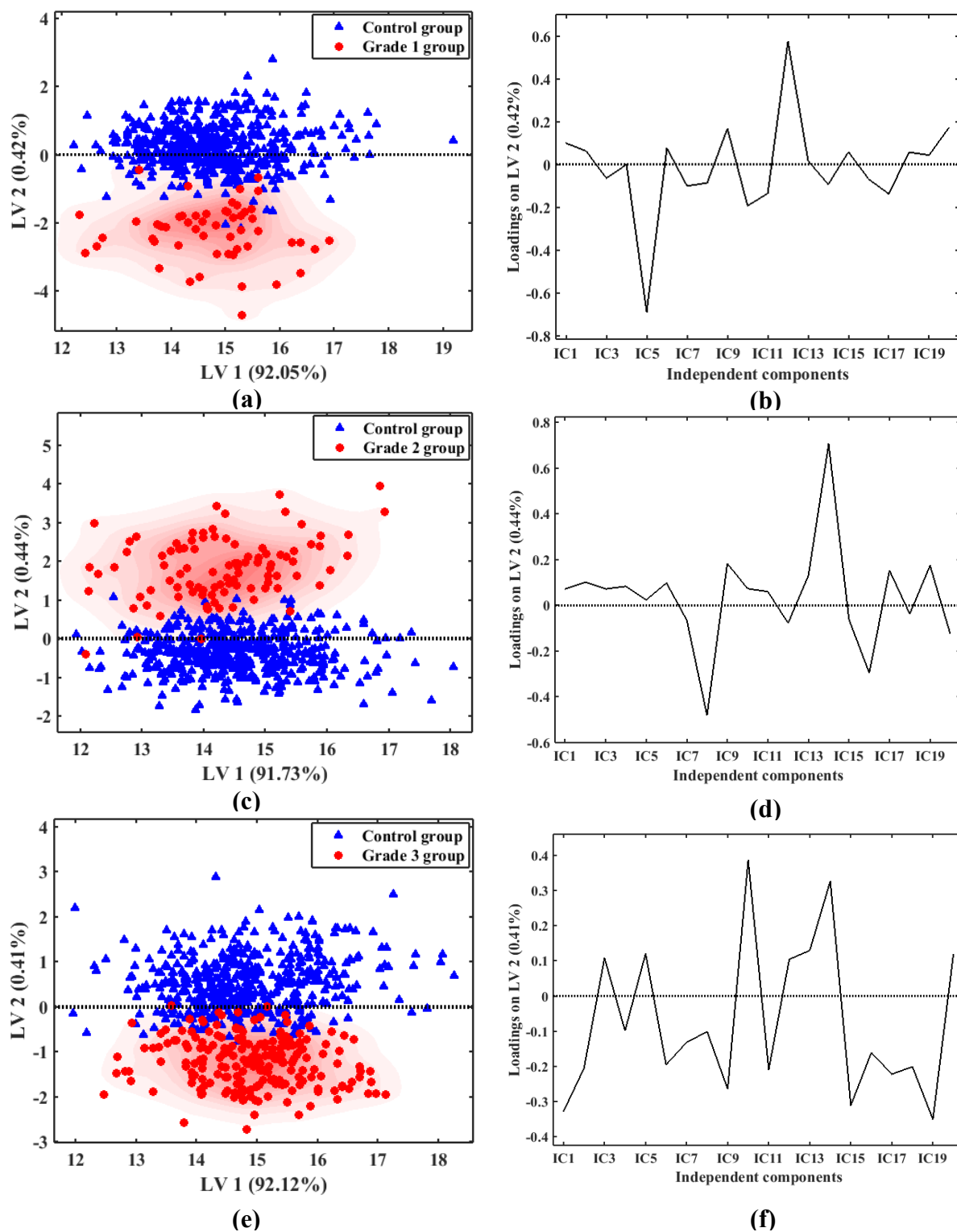


Figure 5.42 The ICA followed by PLS-DA scatter plots for (a) grade 1, (c) grade 2, and (e) grade 3 spectral datasets of saliva samples from control and breast cancer patients. The independent components associated with loadings are shown in parts (b), (d), and (f), respectively.

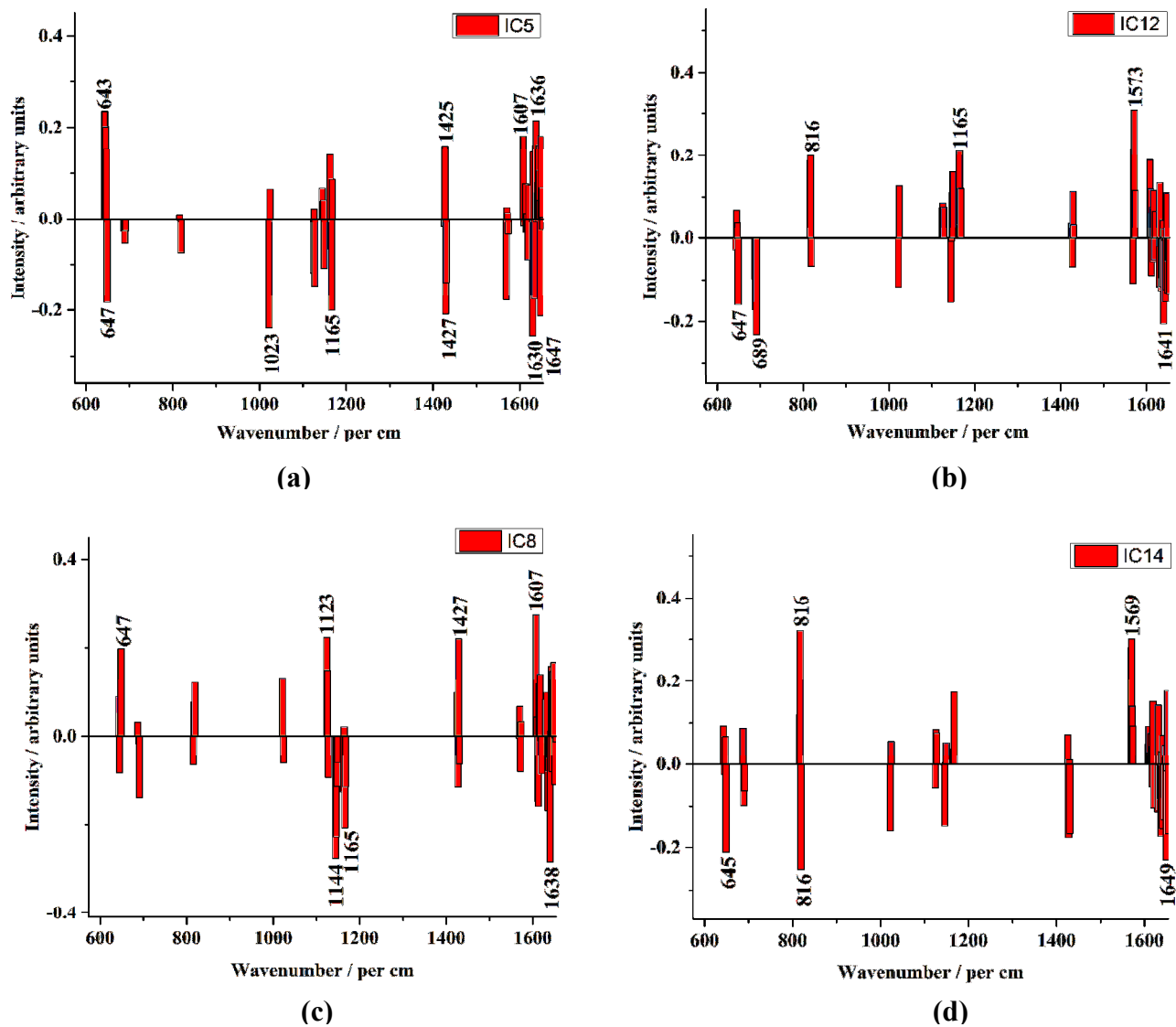


Figure 5.43 The spectral markers for independent components of Raman spectra of saliva samples from healthy volunteers and (a, b) grade 1, and (c, d) grade 2 breast cancer patients.

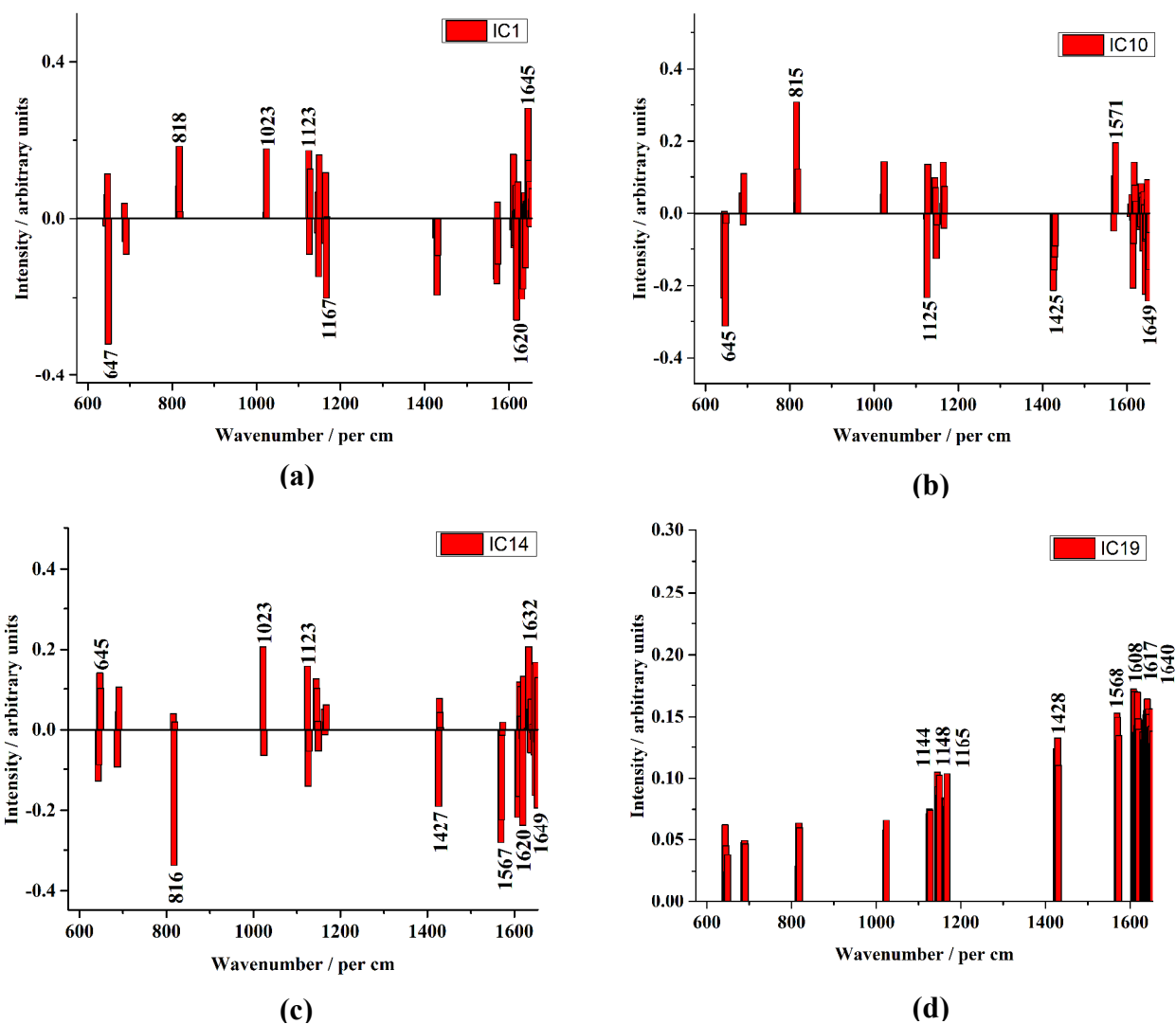


Figure 5.44. The spectral markers for independent components of Raman spectra of saliva samples from healthy volunteers and (a, d) grade 3 breast cancer patients.

led to lower diagnostic sensitivity and specificities. Thus, a different approach which involved use of ICA combined with mahalanobis multidimensional metrics (MDS) and the potential functions (kernel density estimators) was adopted. The mathematical basis of the Mahalanobis distance calculation is well known (McLachlan, 1999), and widely used for spectral discrimination. In addition to providing spectral discrimination, it also gives a statistical measure of how well the unknown sample spectrum matches or does not match (Chowdary *et al.*, 2006). Thus it is a statistical measure of proximity of two spectra.

Table 5.32 Diagnostic results of ICA followed by PLS-DA on the Raman spectra of saliva from healthy (controls) and breast cancer patients

| Disease status | Diagnosis | Cases | | | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|----------------|---------------|---------------|----------|--------------|-------|--------------|-----------------|-----------------|
| | | Breast cancer | Controls | Not assigned | | | | |
| Grade-1 | Breast cancer | 48 | 4 | 2 | 54 | 95% | 89% | 95% |
| | Controls | 13 | 480 | 11 | 504 | | | |
| Grade-2 | Breast cancer | 92 | 3 | 2 | 97 | 95% | 95% | 95% |
| | Controls | 8 | 450 | 16 | 474 | | | |
| Grade-3 | Breast cancer | 191 | 9 | 7 | 207 | 92% | 92% | 92% |
| | Controls | 18 | 431 | 20 | 469 | | | |

The results of ICA-MDS on Raman spectra of saliva samples were fed to kernel density estimators (potential function) algorithm. In the present study, a percentile threshold of 95% was selected, and the sample was classified in a specific class space if its potential was higher than the potential class threshold (Forina *et al.*, 1991). We observed that a continuous Gaussian kernel, smoothing parameters = 0.7-1.2, 9 principal components, and 10-fold cross-validation classification yielded the best model optimization.

Figure 5.45 (a-c) shows the potential function scatter plots of the Raman spectral data of the saliva samples of breast cancer patients of all different stages versus healthy (control) ones. Figure 5.45 (a) shows reasonably good differentiation of the Raman spectral data of healthy (control) and breast cancer patients. The grade 2 and grade 3 samples are well differentiated from controls, although it may be considered that the grade 2 samples are more differentiated in comparison to grade 3 samples (Figure 5.45 (b-c)). The large overlap of the healthy (control) and stage clusters indicates that, although they are clinically distinguished by the extent of the disease progression, they are spectrally, and therefore biochemically partially similar.

A diagnostic sensitivity of 96%, 98%, and 94%, with a specificity of 99%, 98%, and 95%, were achieved for the breast cancer patients and the healthy volunteers, respectively (Table 5.33), which are comparably better than diagnostic sensitivity of 89%, 95%, and 92%, with a specificity of 95%, 95% and 92%, respectively, achieved with ICA followed by PLS-DA (Table 5.32). These results confirm the outstanding diagnostic accuracy of the ICA-MDS-kernel density estimators - based diagnostic algorithm for breast cancer detection.

Table 5.33 Diagnostic results of ICA followed by MDS and kernel density estimators (potential function analysis) on the Raman spectra of saliva from healthy volunteers (controls) and breast cancer patients

| Disease status | Diagnosis | Cases | | | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|----------------|---------------|---------------|----------|--------------|-------|--------------|-----------------|-----------------|
| | | Breast cancer | Controls | Not assigned | | | | |
| Grade-1 | Breast cancer | 52 | 2 | 0 | 54 | 99 | 96 | 99 |
| | Controls | 2 | 500 | 2 | 504 | | | |
| Grade-2 | Breast cancer | 95 | 1 | 1 | 97 | 98 | 98 | 98 |
| | Controls | 5 | 466 | 3 | 474 | | | |
| Grade-3 | Breast cancer | 194 | 7 | 6 | 207 | 95 | 94 | 95 |
| | Control s | 10 | 445 | 4 | 469 | | | |

5.2.2.5 Multivariate statistical analysis of Support Vector Machine (SVM) and Backpropagation neural networks (BPNN) for breast cancer diagnostics in saliva

For classification of saliva spectra from healthy and breast cancer patients, a non-linear classifier were required for analyzing spectral matrices collected at spectral markers of 643-647, 687-689, 816-818, 1022-1024, 1125-1128, 1145-1148, 11164-1166, 1427-1430, 1570-1572, 1609-1619, 11630-1657, and 1753-1756 cm^{-1} . SVM is one of the best classifiers since it finds the hyperplane which maximizes the separating margin between classes (Dehghan *et al.*, 2008). Nonlinear SVM classifier is obtained by first using a nonlinear operators to map the input pattern into a higher dimensional space (Dehghan *et al.*, 2008). We aimed to employ SVMs with Gaussian Radial Basis Function (RBF) kernel as component nonlinear classifier in analysis. For comparison, separate analysis was performed using a linear kernel function. The reason why RBF-SVM component classifiers are favored lies in the fact that these classifiers often have larger diversity than those component classifiers which may be considered accurate, suggesting they may lead to a better generalization performance (Xuchun *et al.*, 2007). Moreover, better results using RBF kernel can be explained due the fact it has less numerical difficulties (Hsu *et al.*, 2016). The cross-validation can be used to determine model parameter in order to avoid exhaustive parameter search by approximations or heuristics (Hertzmann *et al.*, 2012). In the present study, the $k = 10$ -fold cross validation was used to minimize the bias associated with random sampling of training and

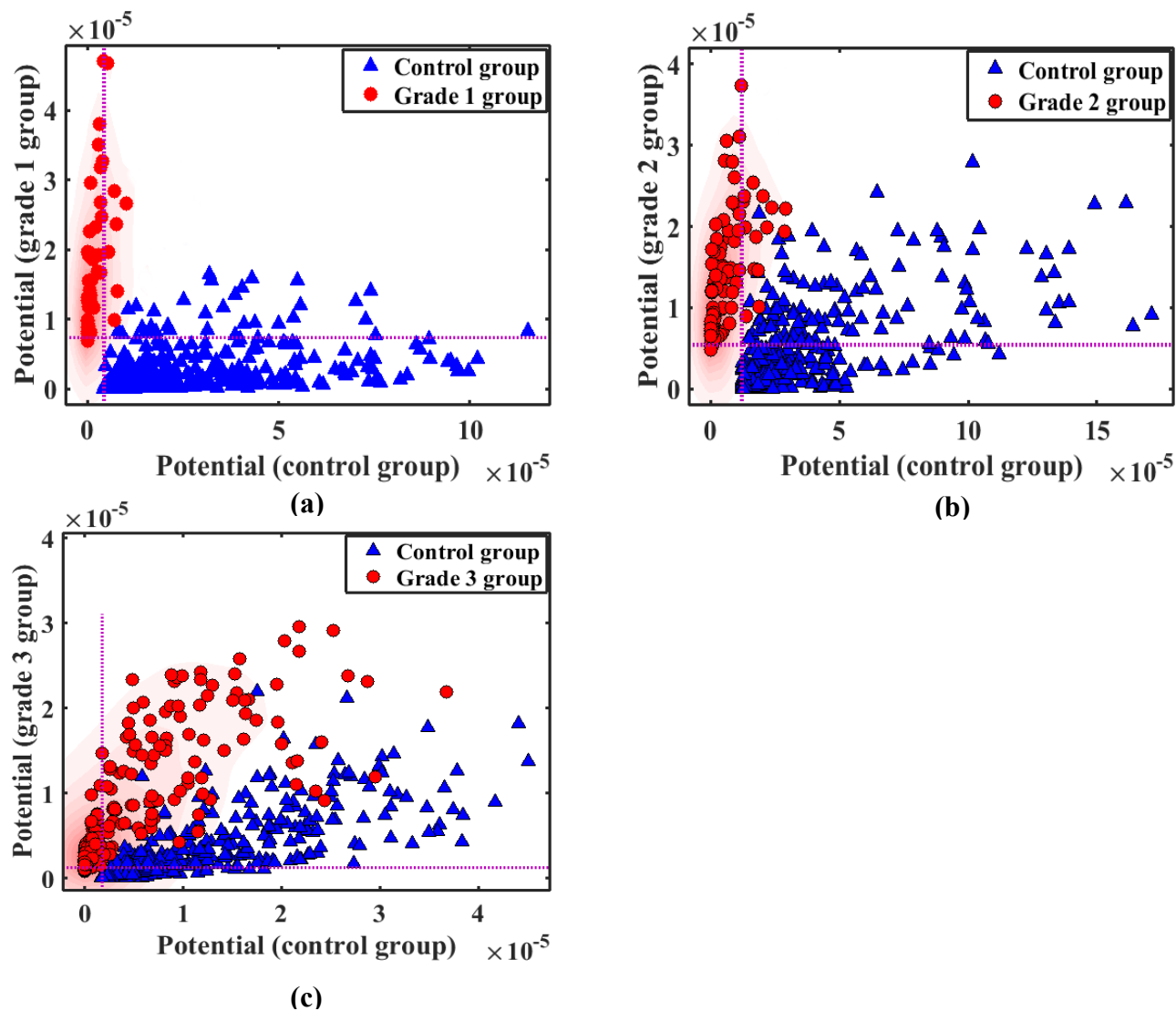


Figure 5.45 The diagnostic results of ICA followed by multidimensional scaling and kernel density estimators for (a) grade 1 breast cancer, (b) grade 2 breast cancer, and (c) grade 3 breast cancer, based on Raman spectra of saliva from healthy volunteers and breast cancer patients.

test data samples in comparing predictive accuracy of two or more methods (Delen *et al.*, 2005). That is, we divided the dataset into 10 mutually exclusive partitions (folds) using a stratified sampling technique. Then, we used 9 of 10 folds for training and the 10th for the testing. We repeated this process for 10 times so that each and every data point would be used as part of the training and testing datasets. The accuracy measure for the model was calculated by averaging the 10 models performance numbers. We repeated this process for each of the two prediction models. This provided us with a less biased prediction performance measures to compare the two models.

The optimal values for principal components (PCs), error penalty (cost, ‘c’) and gamma ‘g’ (kernel parameter), calculated using 10 fold-cross validation, came out to be 10, 100, and 0.57-1.13, respectively as shown in Table 5.34. Table 5.35 and Table 5.36 show the accuracy success achieved by the training and predictive SVM models using different kernel methods, respectively.

Table 5.34. SVM models characteristics for diagnostic analysis on the Raman spectra of saliva samples from healthy (controls) and breast cancer patients

| SVM optimal characteristic | | | | | |
|----------------------------|----------|------|-----|-------------------|-----------------|
| Disease status | Function | Cost | PCs | Kernel parameters | Support vectors |
| Grade-1 | Kernel | 100 | 10 | - | 94 |
| | RBF | 100 | 10 | 0.8 | 97 |
| Grade-2 | Kernel | 100 | 10 | - | 84 |
| | RBF | 100 | 10 | 0.57 | 176 |
| Grade-3 | Kernel | 100 | 10 | - | 172 |
| | RBF | 100 | 10 | 1.13 | 223 |

Initial analysis based on the training model (Table 5.35) suggested that RBF kernel yielded the best performance in terms of accuracy and sensitivity, proving it can be an useful machine learning technique for diagnosis of breast cancer. Other works (Singla *et al.*, 2011; Yang *et al.*, 2007; Lyng *et al.*, 2019) have shown the RBF kernel to be a good classifier for eye event detection, detection and classification of microcalcifications, and classification of benign lesions and breast cancer, respectively. Figure 5.46 shows the Raman spectra of saliva from healthy volunteers (controls) was well separated from Raman spectra of saliva from breast cancer patients, using RBF-SVM training model.

Table 5.35. Diagnostic results of linear-SVM and RBF-SVM models on the Raman spectra of saliva from healthy volunteers (controls) and breast cancer patients

| Disease status | Function | Diagnosis | Cases | | | Accuracy | Sensitivity | Specificity |
|----------------|----------|---------------|---------------|----------|-------|----------|-------------|-------------|
| | | | Breast cancer | Controls | Total | | | |
| Grade-1 | Linear | Breast cancer | 25 | 29 | 54 | 94 | 46 | 100 |
| | | Controls | 2 | 503 | 505 | | | |
| | RBF | Breast cancer | 37 | 17 | 54 | 94 | 69 | 97 |
| | | Controls | 17 | 487 | 505 | | | |
| Grade-2 | Linear | Breast cancer | 75 | 22 | 97 | 94 | 77 | 98 |
| | | Controls | 11 | 463 | 474 | | | |
| | RBF | Breast cancer | 76 | 21 | 97 | 94 | 78 | 98 |
| | | Controls | 10 | 464 | 474 | | | |
| Grade-3 | Linear | Breast cancer | 158 | 49 | 207 | 89 | 76 | 95 |
| | | Controls | 24 | 445 | 469 | | | |
| | RBF | Breast cancer | 190 | 17 | 207 | 96 | 92 | 97 |
| | | Controls | 13 | 456 | 469 | | | |

A similar observation is made from the results of RBF-SVM predictive model (Figure 5.47), though control and diseased scores are largely overlapped which explains low diagnostic accuracy in terms of sensitivity. Moreover, results of predictive model (Table 5.36) demonstrated relative poor performance, with the best performance in terms of sensitivity being 78%. This was most likely due to the low number of patient samples or / and the complexity of the model; particularly, the increase in the size or the number of parameters in the machine learning model, which could have contributed to overfitting. Thus, artificial neural networks (ANN) was chosen for chemometric analysis.

Artificial neural networks (ANNs) are commonly known as biologically inspired, highly sophisticated analytical techniques, capable of modeling extremely complex non-linear functions (Delen *et al.*, 2005). In the present study, we used a multi-layer perceptron (MLP)-based back-propagation (a supervised learning algorithm) technique, a known powerful function approximator for prediction and classification problems (Delen *et al.*, 2005). First, we optimized the model by employing different iterations with varying number of hidden layers and hidden nodes.

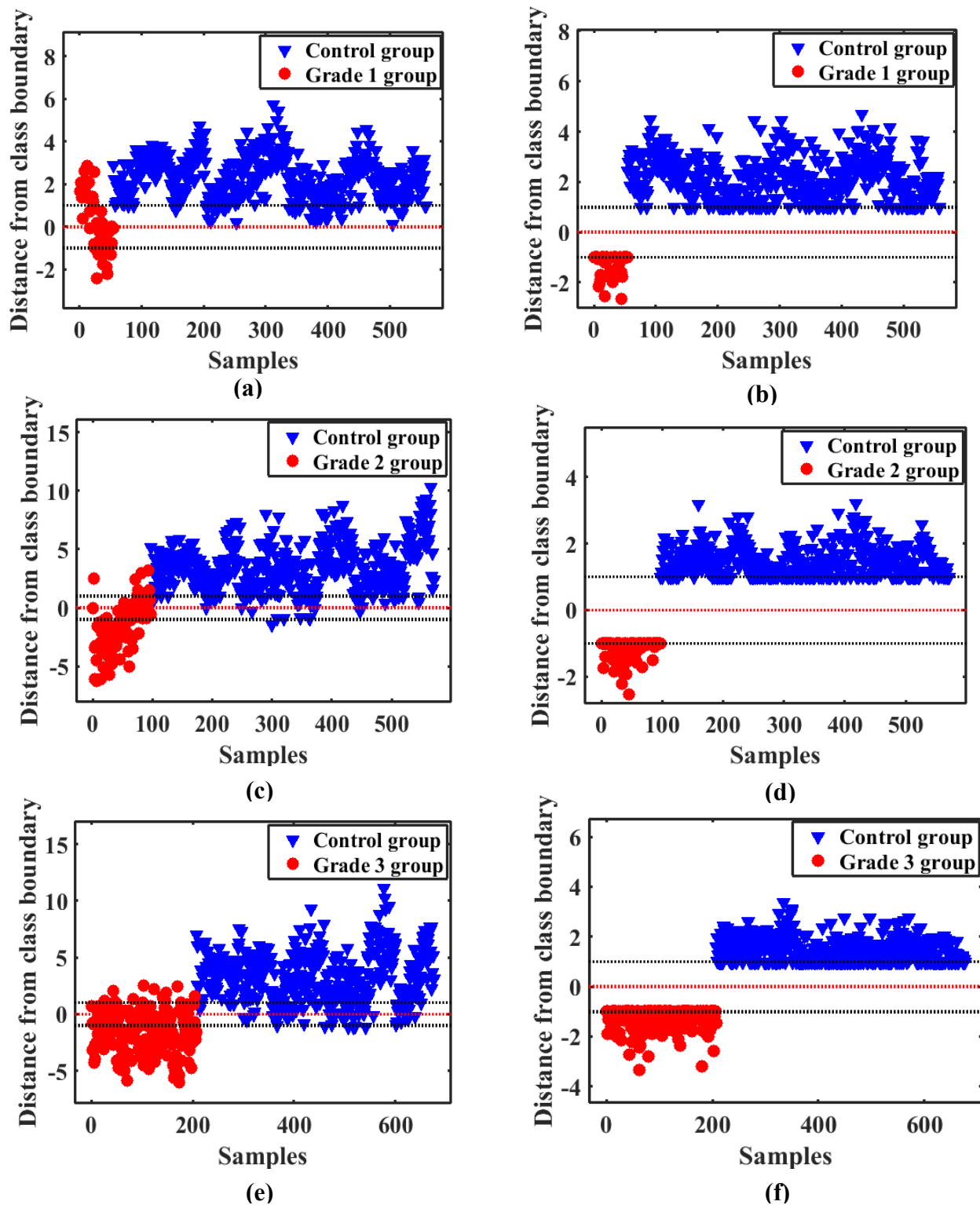


Figure 5.46 The SVM training models for breast cancer detection for (a, b) grade 1 breast cancer, (c, d) grade 2 breast cancer, and (e, f) grade 3 breast cancer. Parts (a), (c) and (e) are linear-SVM training models; (b), (d) and (f) are RBF-SVM training models.

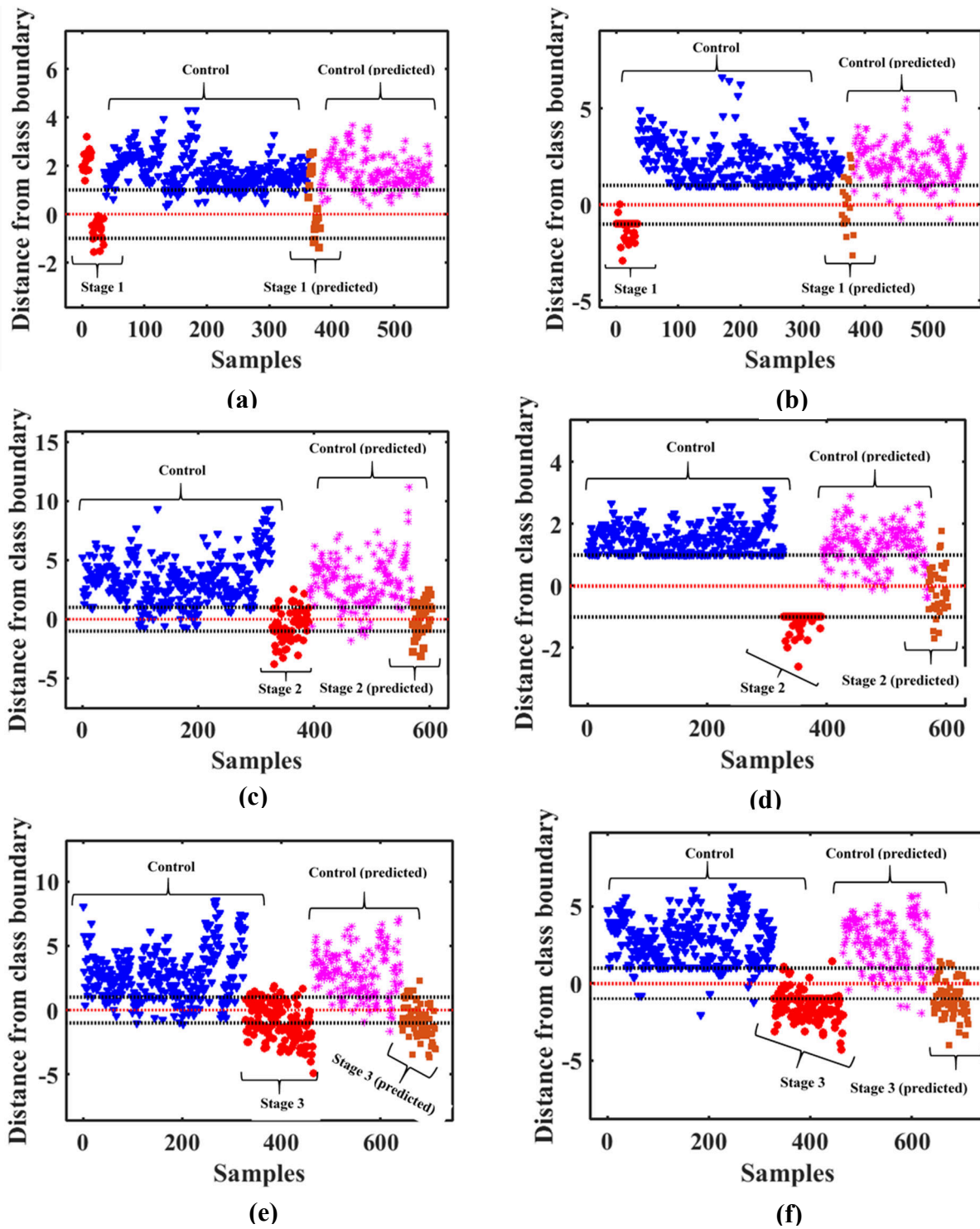


Figure 5.47 SVM prediction models of breast cancer detection for (a, b) grade 1 breast cancer, (c, d) grade 2 breast cancer, and (e, f) grade 3 breast cancer. Parts (a), (c) and (e) are linear-SVM predictive models; (b), (d) and (f) are RBF-SVM predictive models.

Table 5.36. Diagnostic results of linear-SVM and RBF-SVM predictor models on the Raman spectra of saliva samples from healthy (controls) and breast cancer patients

| Disease status | Function | Diagnosis | Cases | | | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|----------------|----------|---------------|---------------|----------|-----|-------|--------------|-----------------|-----------------|
| | | | Breast cancer | Controls | | | | | |
| Grade-1 | Linear | Breast cancer | 10 | 9 | 19 | 95 | 53 | 100 | |
| | | Controls | 0 | 178 | 178 | | | | |
| | RBF | Breast cancer | 11 | 8 | 19 | 92 | 58 | 97 | |
| | | Controls | 5 | 173 | 178 | | | | |
| Grade-2 | Linear | Breast cancer | 17 | 16 | 33 | 88 | 52 | 94 | |
| | | Controls | 10 | 168 | 178 | | | | |
| | RBF | Breast cancer | 19 | 14 | 33 | 91 | 58 | 97 | |
| | | Controls | 6 | 172 | 178 | | | | |
| Grade-3 | Linear | Breast cancer | 50 | 19 | 69 | 89 | 72 | 96 | |
| | | Controls | 7 | 171 | 178 | | | | |
| | RBF | Breast cancer | 54 | 15 | 65 | 90 | 78 | 94 | |
| | | Controls | 10 | 168 | 178 | | | | |

For each training and predicting algorithm, the network architecture was varied for 2 hidden layers, with the nodes varied from 5 to 100 at increments of 5 in each hidden layer. Exactly 2 layers, neurons per layer =10, learning rate=0.01, alpha=0.5, and iterations=1000 training weights yield the best diagnostic accuracy in classifying early breast malignancy (stage 1 cancer) whereas 2 layers, neurons per layer =15, learning rate=0.01, alpha=0.3, and iterations=1000 training weights performed well in training and predicting middle (stage 2 cancer) and late (stage 3 cancer) breast malignancies.

Figure 5.48 shows separation of Raman spectra of saliva from the healthy volunteers (controls) and the breast cancer patients. It is evident there was reasonable separation of scores during MLP network model training and prediction analysis. As observed in Tables 5.37 and 5.38, the diagnostic results of BPNN training and predictive models were better than the linear-SVM and RBF-SVM training and predictive models (Tables 5.35, 5.36). The best accuracies and sensitivities that we could obtain were above 90% and 80%, respectively, for all stages of breast cancer under consideration, which was higher than for SVM models we developed. We believe

the performance of MLP can be improved further when applied on a larger data set with more features.

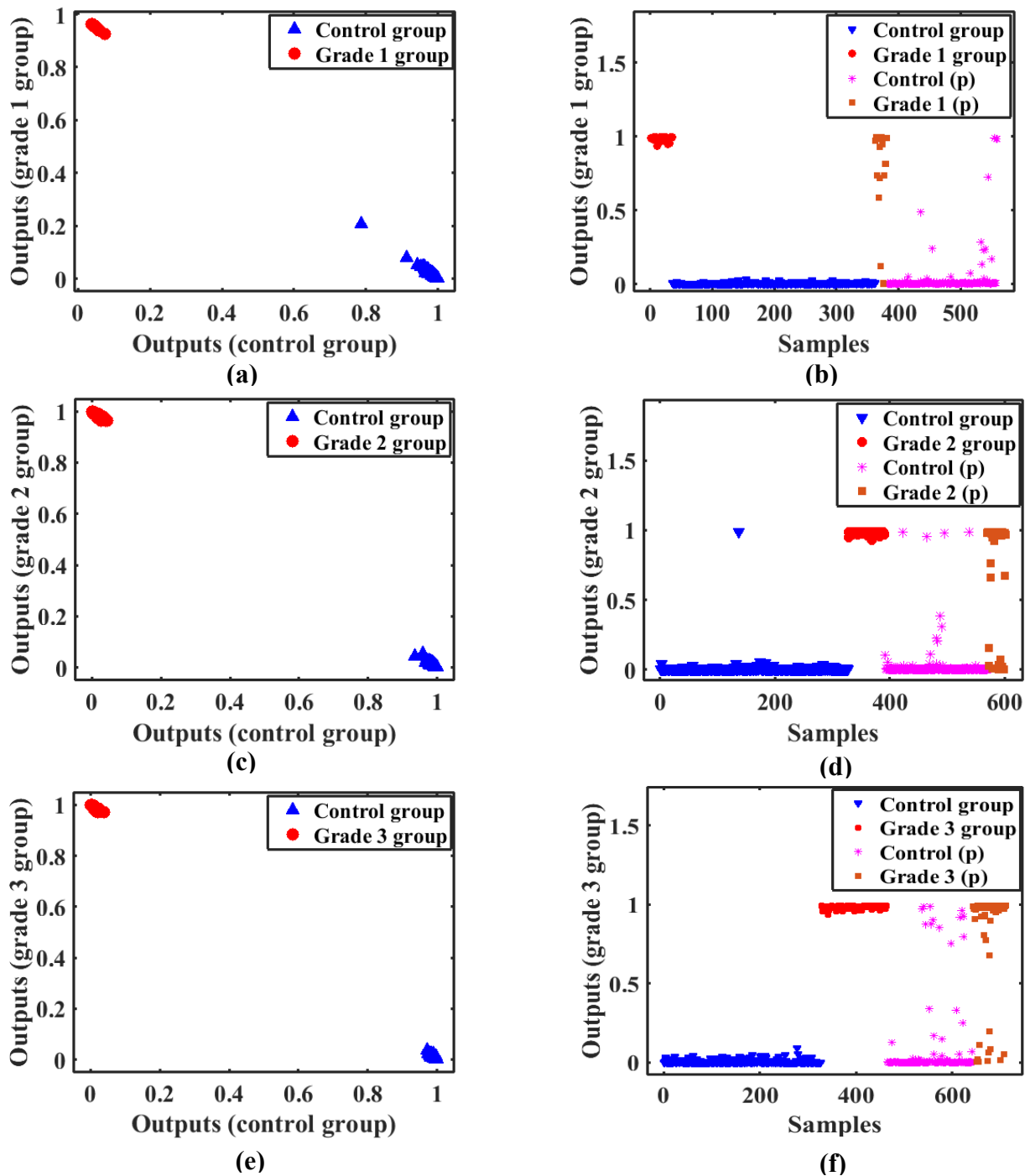


Figure 5.48 The BPNN training (a, c, e) and predictor (b, d, f) models for detecting (a, b) grade 1 breast cancer, (c, d) grade 2 breast cancer, and (e, f) grade 3 breast cancer. Abbreviation: P - predicted samples (scores).

Table 5.37 Diagnostic results of BPNN training model on the Raman spectra of saliva samples from healthy volunteers (controls) and breast cancer patients

| Cases | | | | | | | | |
|----------------|---------------|---------------|----------|--------------|-------|--------------|-----------------|-----------------|
| Disease status | Diagnosis | Breast cancer | Controls | Not assigned | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
| Grade-1 | Breast cancer | 49 | 5 | 0 | 54 | 96 | 91 | 97 |
| | Controls | 19 | 480 | 6 | 505 | | | |
| Grade-2 | Breast cancer | 84 | 12 | 1 | 97 | 92 | 88 | 93 |
| | Controls | 33 | 435 | 6 | 474 | | | |
| Grade-3 | Breast cancer | 180 | 25 | 2 | 207 | 91 | 88 | 93 |
| | Controls | 33 | 430 | 6 | 469 | | | |

Table 5.38. Diagnostic results of BPNN predictor model on the Raman spectra of saliva samples from healthy volunteers (controls) and breast cancer patients

| Cases | | | | | | | | |
|----------------|---------------|---------------|----------|--------------|-------|--------------|-----------------|-----------------|
| Disease status | Diagnosis | Breast cancer | Controls | Not assigned | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
| Grade-1 | Breast cancer | 15 | 3 | 1 | 19 | 96 | 83 | 97 |
| | Controls | 5 | 169 | 4 | 178 | | | |
| Grade-2 | Breast cancer | 28 | 5 | 0 | 33 | 92 | 84 | 98 |
| | Controls | 4 | 174 | 1 | 179 | | | |
| Grade-3 | Breast cancer | 58 | 11 | 0 | 69 | 91 | 84 | 93 |
| | Controls | 12 | 166 | 0 | 178 | | | |

5.3 Raman spectroscopic characterization of whole blood and saliva for leukemia diagnostics

5.3.1 Analysis of prominent biochemical alterations in whole blood and saliva spectra

For a better comparison of Raman spectral shapes in the analysis, Raman spectra for healthy volunteers (controls) and leukemia groups' samples were plotted alongside their respective difference spectra. To better comprehend the molecular basis for the observed Raman spectra, the tentative assignments of the Raman bands were performed according to the known literature (Movasaghi *et al.*, 2007; Rehman *et al.*, 2013; Gelder *et al.*, 2007). The average Raman spectra of whole blood and saliva samples of the leukemia and control groups are shown in Figures 5.49 (a) and (b), respectively. Respective difference spectra are shown in Figure 5.50 (a) and (b).

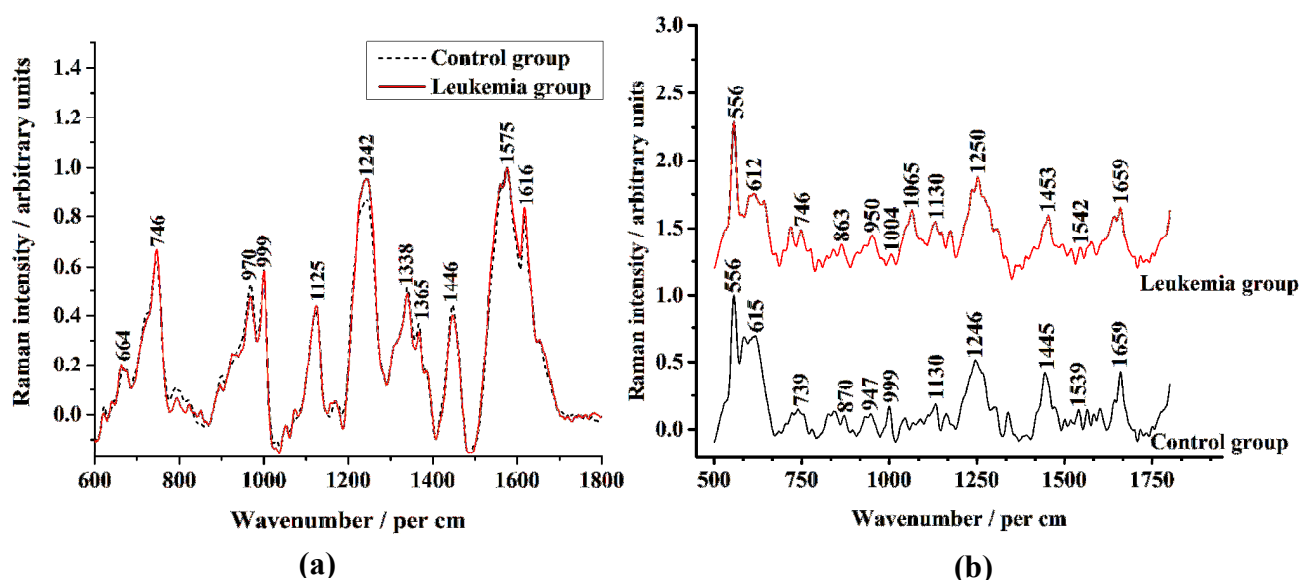


Figure 5.49 Mean normalized spectra of (a) whole blood and (b) saliva for healthy volunteers / controls ($n = 12$) and leukemia ($n = 9$) patients. The spectra in (b) have been linearly offset for clarity.

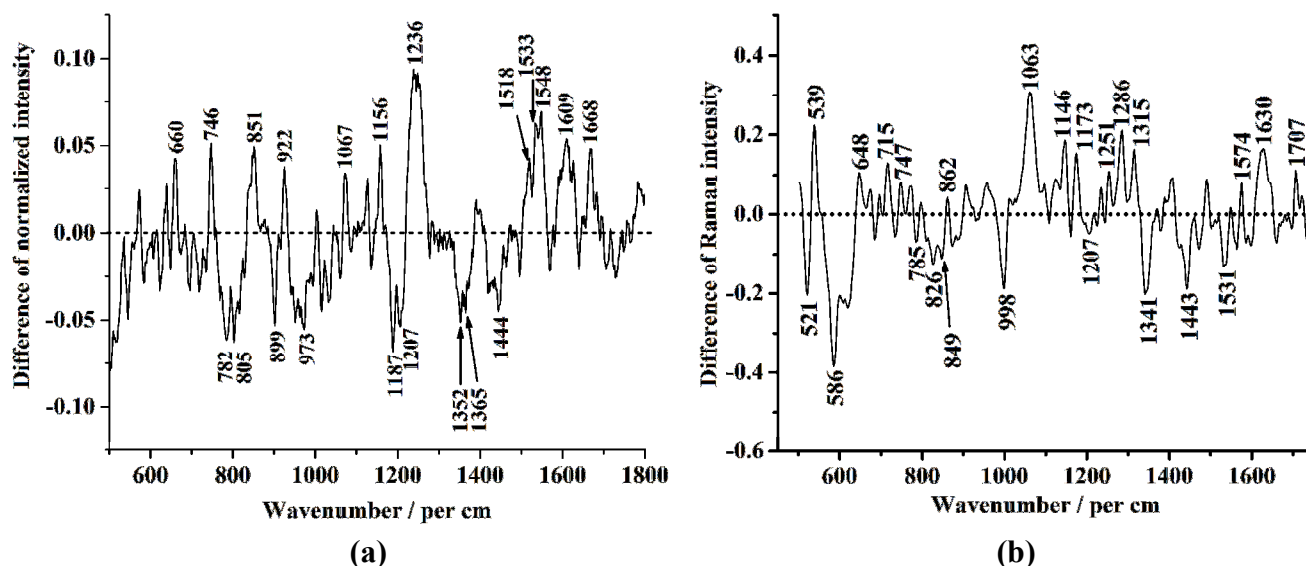


Figure 5.50 The difference spectrum for (a) blood spectra, and (b) saliva spectra from healthy volunteers (controls) and leukemia patients.

These spectra (Figure 5.49 (a), (b)) display Raman bands associated with proteins, lipids, and nucleic acids as the main constituents of the cellular components, which agrees with other related works (Happillon *et al.*, 2015). Examination of blood spectra (Figure 5.49 (a)) show the primary Raman bands featured at 664 cm^{-1} , 746 cm^{-1} , 970 cm^{-1} , 999 cm^{-1} , 1125 cm^{-1} , 1242 cm^{-1} , 1338 cm^{-1} , 1365 cm^{-1} , 1446 cm^{-1} , 1575 cm^{-1} , and 1616 cm^{-1} , which correspond to other findings (Vanna *et al.*, 2014; Sheng *et al.*, 2013). It can be observed (Figure 5.49 (b)), that the Raman spectra of saliva samples exhibited similar peaks at $739 / 746\text{ cm}^{-1}$, $999 / 1004\text{ cm}^{-1}$, 1130 cm^{-1} , $1246 / 1250\text{ cm}^{-1}$ and $1445 / 1453\text{ cm}^{-1}$. Moreover, there are notable red shifts and blue shifts of various Raman bands in spectra of saliva samples. For instance, the bands at 615 cm^{-1} and 870 cm^{-1} in Raman spectra of saliva samples of healthy volunteers can be viewed to have red shifted to 612 cm^{-1} and 863 cm^{-1} in Raman spectra of saliva samples of diseased patients. In contrast the bands at 739 cm^{-1} , 947 cm^{-1} , 999 cm^{-1} , 1246 cm^{-1} , 1445 cm^{-1} and 1539 cm^{-1} in Raman spectra of saliva samples of healthy volunteers can be viewed to have blue shifted to 746 cm^{-1} , 950 cm^{-1} , 1004 cm^{-1} , 1250 cm^{-1} , 1453 cm^{-1} and 1542 cm^{-1} in Raman spectra of saliva samples of diseased patients.

Based on available literature (Movasaghi *et al.*, 2007; Rehman *et al.*, 2013; Gelder *et al.*, 2007), it can be observed in Figure 5.49 and Figure 5.50 that peak intensities and band shifts in spectra of whole blood and saliva can be mainly attributed to biochemical changes due to nucleic

acids (556, 715, 742, 746, 782, 785, 805, 826, 862, 867, 1173, 1187, 1207, 1286, 1315, 1338, 1341, 1352, 1365, 1518, 1575, 1609 cm^{-1}), proteins (539, 568, 648, 660, 664, 849, 851, 922, 948, 970, 990, 1001, 1156, 1236, 1242, 1248, 1251, 1443, 1449, 1531, 1533, 1540, 1548, 1616, 1630, 1668, 1707 cm^{-1}) and lipids (521, 613, 1063-1067, 1125, 1444-1449 cm^{-1}). With regard to blood spectra, nucleic acids, proteins and lipids biochemical alterations in normal and leukemia samples are evident at 664 cm^{-1} (C-S stretch mode of collagen), 746 cm^{-1} (thymine), 970 cm^{-1} (phosphate monoester groups of phosphorylated proteins and cellular nucleic acids), 999 cm^{-1} (phenylalanine), 1125 cm^{-1} (C-C skeletal backbone of lipids, C-N stretching of proteins), 1242 cm^{-1} (amide III), 1338 cm^{-1} ($\text{CH}_2 / \text{CH}_3$ wagging, twisting and /or bending mode of collagens and lipids), 1365 cm^{-1} (guanine, tryptophan), 1446 cm^{-1} (CH_2 bending mode of proteins and lipids), 1575 cm^{-1} (ring breathing modes in the DNA bases of guanine, adenine) and 1616 cm^{-1} (C=C stretching mode of tyrosine and tryptophan). Moreover, comparisons of the Raman intensities of the six prominent Raman peaks of blood spectra (664 cm^{-1} , 746 cm^{-1} , 970 cm^{-1} , 1242 cm^{-1} , 1446 cm^{-1} , and 1616 cm^{-1}) showed significant differences ($p < 0.05$; Student's *t*-test).

Figure 5.50 (a) shows the corresponding difference spectra of blood samples, revealing the significant Raman spectral changes, such as Raman peak intensities, positions, and spectral shoulder bands, specifically in the spectral ranges of 660 cm^{-1} (C-S stretching mode of cystine (collagen type I)), 746 cm^{-1} (thymine), 782 cm^{-1} (thymine, cytosine, uracil), 805 cm^{-1} (uracil), 851 cm^{-1} (proline and tyrosine ring breathing), 899 cm^{-1} (saccharides), 922 cm^{-1} (C-C stretch of proline), 973 cm^{-1} (C-C backbone (collagen assignment)), 1067 cm^{-1} (proline (collagen assignment)), 1156 cm^{-1} (C-C, C-N stretching (protein)), 1187 cm^{-1} (cytosine, guanine, adenine), 1207 cm^{-1} (proline, tyrosine), 1236 cm^{-1} (amide III), 1352 cm^{-1} (guanine), 1365 cm^{-1} (tryptophan), 1444 cm^{-1} (CH_2CH_3 bending modes of collagen and phospholipids) and 1500-1700 cm^{-1} (amides, DNA bases, lipids). Compared with the control group samples spectra, the leukemia group spectra exhibited higher intensities at 660 cm^{-1} (C-S stretching mode of cystine (collagen type I)), 746 cm^{-1} (thymine), 851 cm^{-1} (proline and tyrosine ring breathing), 922 cm^{-1} (C-C stretch of proline), 1067 cm^{-1} (proline (collagen assignment)), 1156 cm^{-1} (C-C, C-N stretching (protein)), 1236 cm^{-1} (amide III), 1518 cm^{-1} (β – carotene accumulation), 1533 cm^{-1} (β – carotene accumulation), 1548 cm^{-1} (tryptophan), 1609 cm^{-1} (cytosine) and 1668 cm^{-1} (amide I), but showed much increased signals at 1236 cm^{-1} , 1533 cm^{-1} and 1548 cm^{-1} . In contrast, the control group samples spectra exhibited higher intensities at 782 cm^{-1} (Thymine, cytosine, uracil), 805 cm^{-1} (uracil), 899 cm^{-1} (saccharides), 973 cm^{-1} (phosphate monoester groups of phosphorylated proteins and cellular

nucleic acids), 1187 cm^{-1} (cytosine, guanine, adenine), 1207 cm^{-1} (proline, tyrosine), 1352 cm^{-1} (guanine), 1365 cm^{-1} (tryptophan) and 1444 cm^{-1} (CH_2CH_3 bending modes of collagen and phospholipids).

Literature on Raman spectra of saliva for leukemia diagnostics is scarce, and thus, most bands assignments in the spectra of Figure 5.49 (b) are based on generally known spectral frequencies of the biological tissues (Chandra *et al.*, 2015), (Movasaghi *et al.*, 2007), (Gelder *et al.*, 2007). The Raman spectra of saliva for healthy volunteers (controls) and leukemia groups (Figure 5.50(b)) revealed common peaks (\pm standard deviations) at 556 cm^{-1} (adenine), 613 \pm 0.87 cm^{-1} (cholesterol esters), 742 \pm 2.02 cm^{-1} (thymine), 867 \pm 1.44 cm^{-1} (RNA), 948 \pm 0.87 cm^{-1} (proline, valine, saccharides), 1001 \pm 1.44 cm^{-1} (phenylalanine), 1248 \pm 1.15 cm^{-1} (guanine, cytosine), 1449 \pm 2.31 cm^{-1} (CH_2 bending mode of proteins and lipids) and 1540 \pm 0.87 cm^{-1} (amide II) (Figure 5.50 (b)). The respective difference spectra (Figure 5.50 (b)) indicates biochemical changes due to cholesterol esters (539 cm^{-1}), C-C twisting mode of tyrosine (648 cm^{-1}), C-N deformation of adenine (715 cm^{-1}), thymine (747 cm^{-1}), tyrosine (862 cm^{-1}), skeletal C-C stretch of lipids (1063 cm^{-1}), glycogen (1146 cm^{-1}), cytosine / guanine (1173 cm^{-1}), amide III (1251 cm^{-1}), cytosine (1286 cm^{-1}), guanine (1315 cm^{-1}), guanine / adenine / tryptophan (1574 cm^{-1}), amide I (1630, 1707 cm^{-1}) were intense in saliva spectra of leukemia group samples whereas biochemical changes due to phosphatidylinositol (521 cm^{-1}), glycerol (586 cm^{-1}), adenine (785 cm^{-1}), O-P-O stretch of DNA (826 cm^{-1}), proline / valine / saccharides (849 cm^{-1}), phenylalanine (998 cm^{-1}), phenylalanine / adenine / thymine (1207 cm^{-1}), guanine (1341 cm^{-1}), CH_2 bending mode of proteins and lipids (1443 cm^{-1}), and amide II (1531 cm^{-1}) were intense in saliva spectra of control group samples.

To assess the diagnostic accuracy of leukemia based on whole blood and saliva spectra, spectral differences were further explored in detail by the SVD-PCA multivariate algorithm as described in *Section 4.6*. The LDA diagnostic model coupled with the $k(= 10)$ -fold cross-validation method was subsequently utilized as a diagnostic algorithm. The first six principal components (PCs) were found to be the optimal number of reserved components (Figure 5.51 (a), (b)), as defined by the part minimum of the root mean square error of the cross-validation, accounting for 91.13% and 98.725% of the whole Raman spectral variances in whole blood spectra and saliva spectra, respectively. By analysis of canonical variables distribution (Figure 5.51 (c), (d)), t -tests, and effect sizes, it was observed that PC 2 ($p < 0.05$, Cohen $d = 1.70$) and PC 5 ($p < 0.05$, Cohen $d = 0.91$) were significant for further analysis of blood spectra, whereas PC 2 ($p <$

0.05, Cohen $d = 2.45$) and PC 6 ($p < 0.05$, Cohen $d = 0.43$) were significant for analysis of saliva spectra.

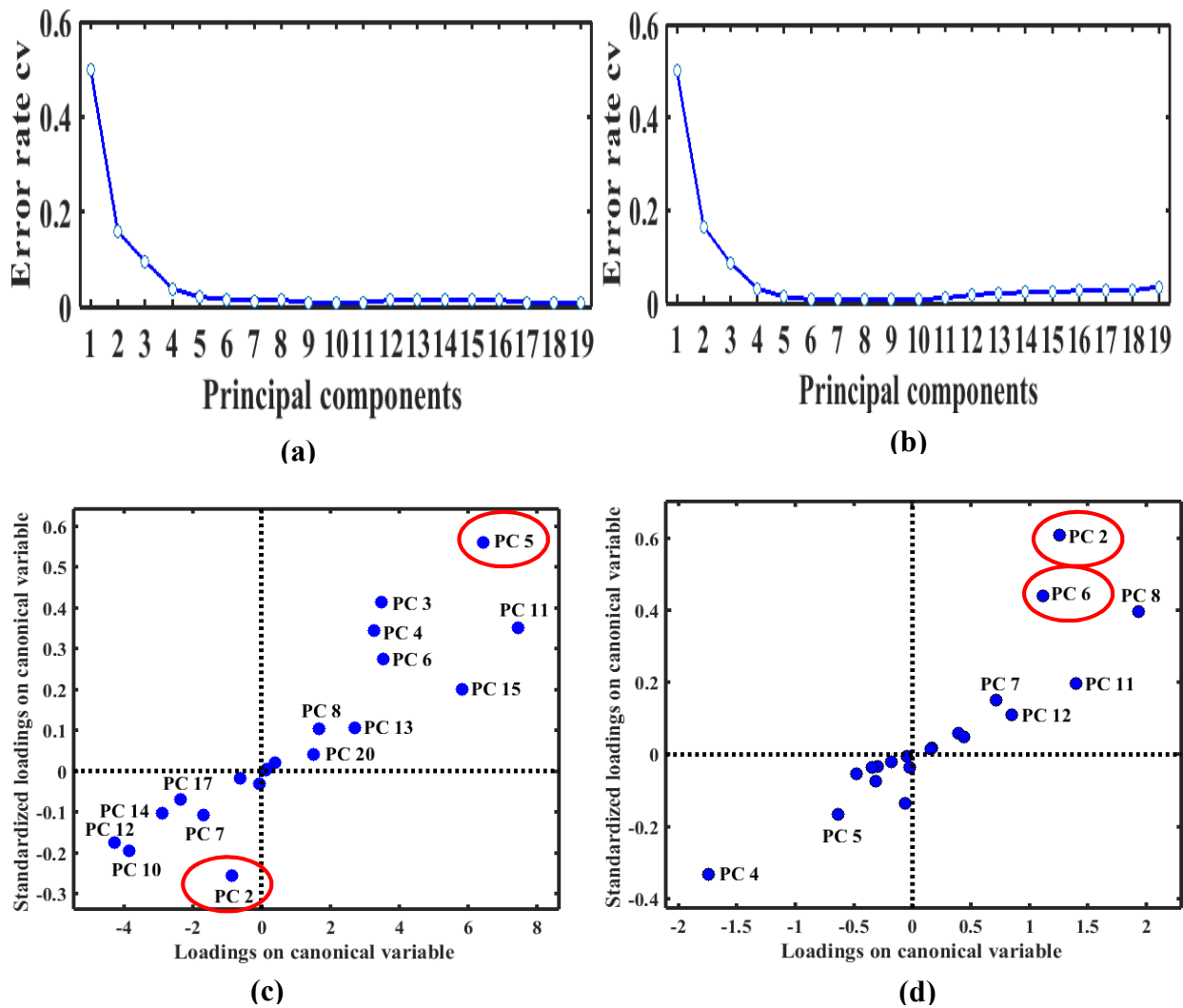


Figure 5.51 The scree plots showing the number of optimal number of principal components (PCs) for (a) blood spectra and (b) saliva spectra, and the canonical variable distributions of the first twenty principal components for (c) blood spectra and (d) saliva spectra. Abbreviations: CV, cross-validation; PC, principal components.

Linear discriminate analysis was utilized to generate a diagnostic algorithm using the first six significant principal components. Figures 5.52 (a, b) shows the scatter plots of each sample according to the first two discriminant functions, with diagnostic lines of LDA clearly indicated.

The loading vectors explaining scores discrimination of blood and saliva spectra are shown in Figures 5.52 (c) and (d), respectively.

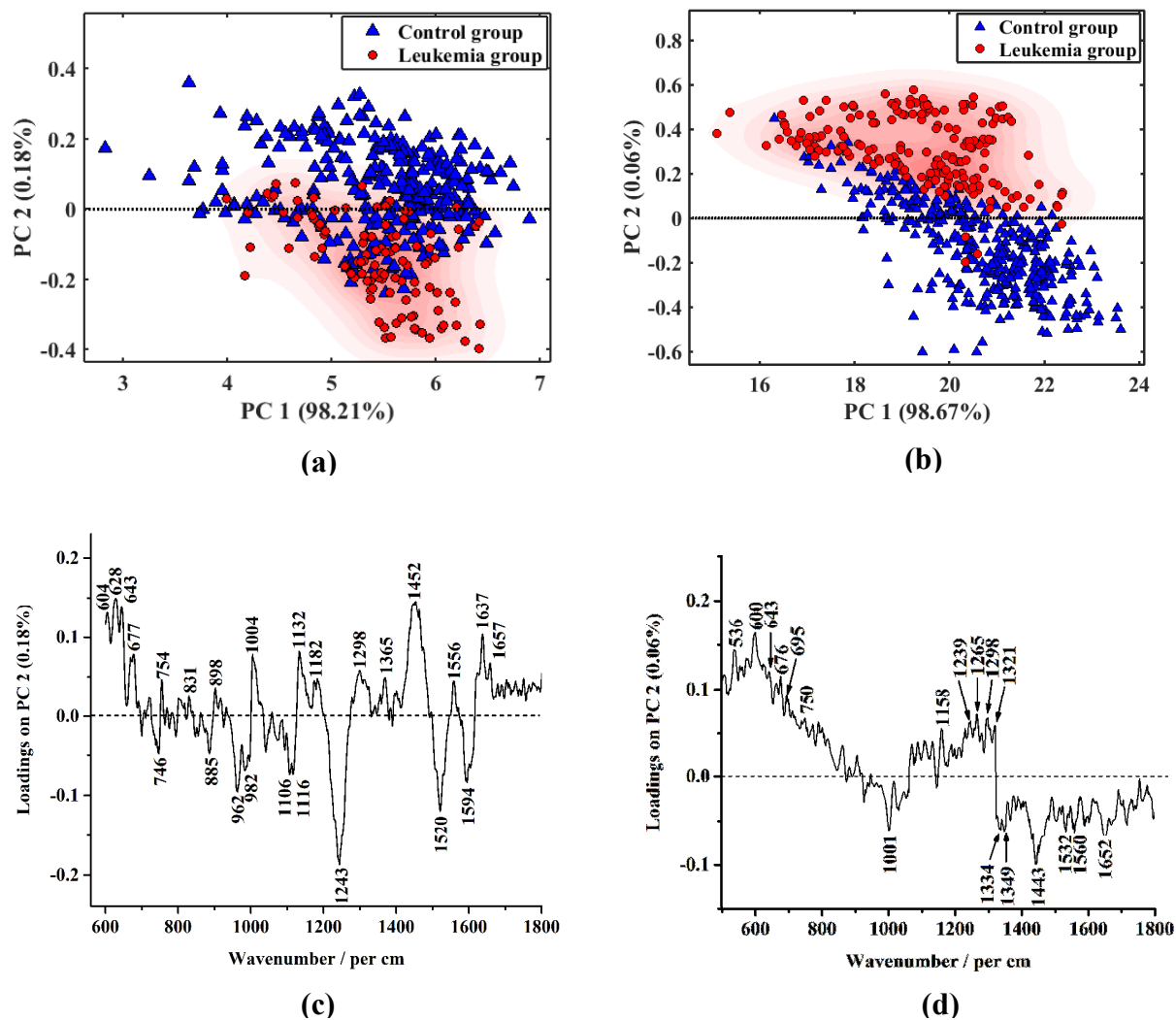


Figure 5.52 Scatter plot of the linear discriminant analysis demonstrating the clustering of (a) whole blood spectra and (b) saliva spectra of healthy volunteers and leukemia patients. The loading vectors explaining the scores discrimination are shown in parts (c) and (d), respectively.

It can be seen (Figure 5.52 (a), (b)) that scores were distributed in two relatively separate areas i.e., control and leukemia groups in spite of some overlap between each other, which indicates that the Raman spectra of the different types of whole blood and saliva samples could be discriminated and classified for leukemia detection. As expected, the loading vectors of the PC 2 in Figure 5.52 (a) and Figure 5.52 (b) are very similar to the difference spectrum of Figure 5.50

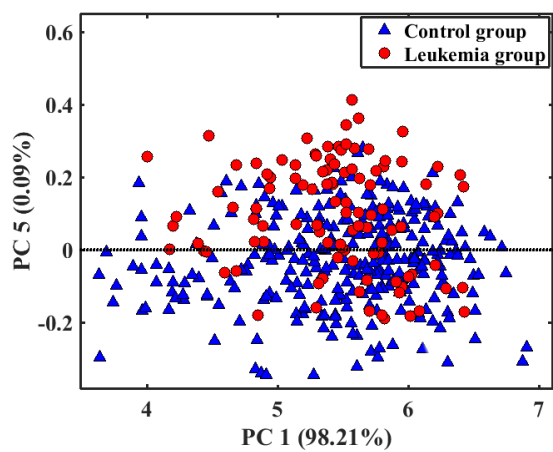
(a) and Figure 5.51 (b), respectively, suggesting that the prominent biochemical differences observed between the spectra of the two groups might be sufficient to tell them apart.

5.3.2 Analysis of trace biochemical alterations of blood and saliva for leukemia diagnosis

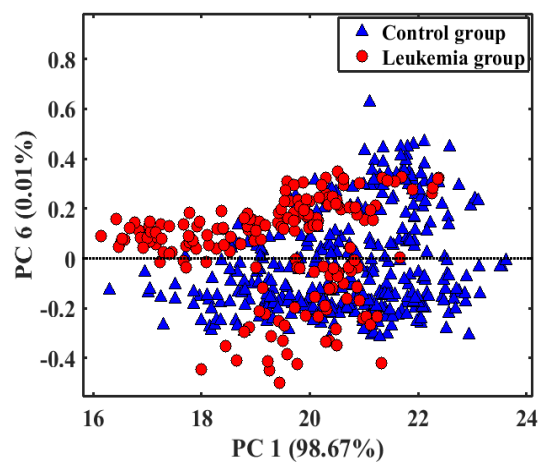
In the present study, subtle but discernible differences in the loading vectors of the Raman spectra of healthy volunteers and leukemia patients were observed, based on scores distribution of the fifth and sixth principal components (PCs 5, 6) as shown in Figure 5.53 (a), (b). The respective loading vectors are shown in Figure 5.53 (c), (d). It is clear that, despite greater overlap of control group and leukemia group scores, the intense loading vectors could be related to the few amount of scores giving rise to the positive and / or negative bands in the loading spectrum.

With regard to blood samples (Figure 5.53 (c)), subtle features of C-S stretching and C-C twisting of proteins- tyrosine at 639 cm^{-1} , amino acids of proline and valine at 923 cm^{-1} , phenylalanine at 999 cm^{-1} , fatty acids at 1130 cm^{-1} , carotenoids at 1158 cm^{-1} , amide III at 1197 cm^{-1} , antisymmetric phosphate stretching vibration at 1227 cm^{-1} , CH_2 twisting modes of lipids at 1301 cm^{-1} , guanines at 1346 cm^{-1} , guanines / tryptophan at 1369 cm^{-1} , and nucleic acids at 1459 cm^{-1} were influential for the assignment of scores into the leukemia group. On the other hand, the contents of cytosine / tyrosine / phenylalanine at 1605 cm^{-1} , amide I at 1697 cm^{-1} , ester groups at 1729 cm^{-1} , and lipids at 1769 cm^{-1} led to a classification of control group spectra.

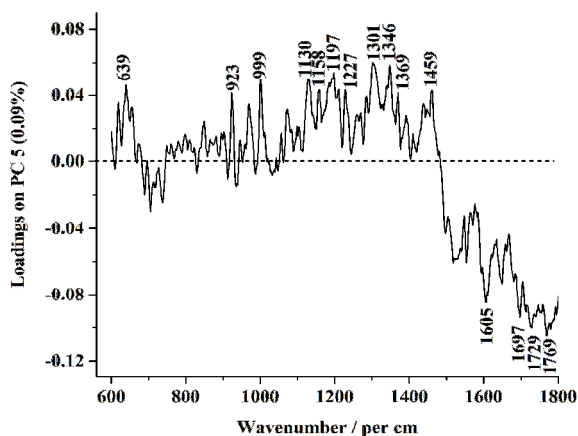
If we consider saliva samples (Figure 5.53 (d)), subtle features of glycerol at 591 cm^{-1} , C-C twisting mode of phenylalanine at 619 cm^{-1} , C-C stretching mode of proline and valine at 935 cm^{-1} , and amide I at 1660 cm^{-1} led to classification of leukemia group spectra whereas the subtle biochemical changes due to C-C twisting mode of tyrosine at 652 cm^{-1} , guanine at 685 cm^{-1} , C-N deformation of nucleic acids at 718 cm^{-1} , thymine / adenine / guanine at 1372 cm^{-1} , and cytosine at 1507 cm^{-1} were influential for the assignment into the control group. In the present study, these subtle markers (loading vectors) represented the weak variance signals (analyte information) significant for leukemia diagnostics, and were therefore chosen for further analysis.



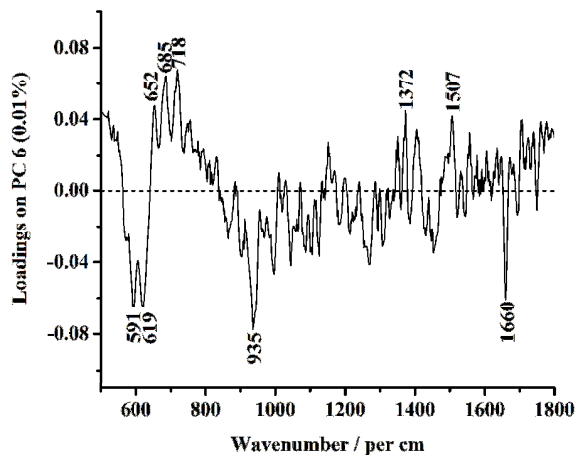
(a)



(b)



(c)



(d)

Figure 5.53 Scores plot for the higher order components (PC 5, 6) of (a) whole blood and (b) saliva Raman spectra (red leukemia samples, blue controls), and (c, d) loading vectors for PC 5 and PC 6, respectively.

5.3.3 Quantitative analysis of trace biomarkers in blood and saliva using partial least-squares regression

The biochemical assignments corresponding to the observed loading vectors (Figure 5.53 (c) (d)) were quantified using the partial least squares (PLS) regression model, as described in Section 4.6.3. The predicted versus measured regression plots in Figure 5.54 and Figure 5.55 show how the PLS model predicted concentration levels for the calibration samples of whole blood and saliva, respectively. The limits of detection for biochemical compounds in simulated whole blood and saliva samples are summarized in Table 5.39 and Table 5.40, respectively. The low root mean-square error of prediction (RMSEP) demonstrates the model had a higher predictive ability (Gontijo *et al.*, 2014), which agrees with the corresponding higher R^2 values (> 0.9). Besides, limits of detection were within the acceptable range of calibration set. LOD values suggested there were adequate analyte concentration present to yield an analytical signal that could be well measured from analytical noise, whereas LOQ demonstrated quantitative results could be obtained with a specified degree of confidence (Taleuzzaman, 2018). Moreover, the accuracy and reliability of the PLS regression model assessed by analyzing concentration levels of a standard simulated blood fluid and saliva spiked with known concentrations of biochemical components demonstrated the calculated biochemical components levels were in agreement with known values in typical standard samples in the range of $\leq 8\%$ and $\leq 3\%$, respectively (Tables 5.41, 5.42).

The relative amounts of biochemical components (in mg / ml) were calculated by fitting the basal spectra in spectral datasets of the spectra markers measured from blood sample (639 cm^{-1} , 923 cm^{-1} , 999 cm^{-1} , 1130 cm^{-1} , 1158 cm^{-1} , 1197 cm^{-1} , 1227 cm^{-1} , 1301 cm^{-1} , 1346 cm^{-1} , 1369 cm^{-1} , 1459 cm^{-1} , 1605 cm^{-1} , 1697 cm^{-1} , 1729 cm^{-1} , and 1769 cm^{-1}) and saliva sample (591 cm^{-1} , 619 cm^{-1} , 652 cm^{-1} , 685 cm^{-1} , 718 cm^{-1} , 935 cm^{-1} , 1372 cm^{-1} , 1507 cm^{-1} , and 1660 cm^{-1}). The determined amounts of biochemical components in whole blood and saliva samples are summarized in Table 5.43 and Table 5.44, respectively. For comparison between the healthy (controls) and diseased patients, the determined concentration levels (in mg / ml) were normalized to their mean value.

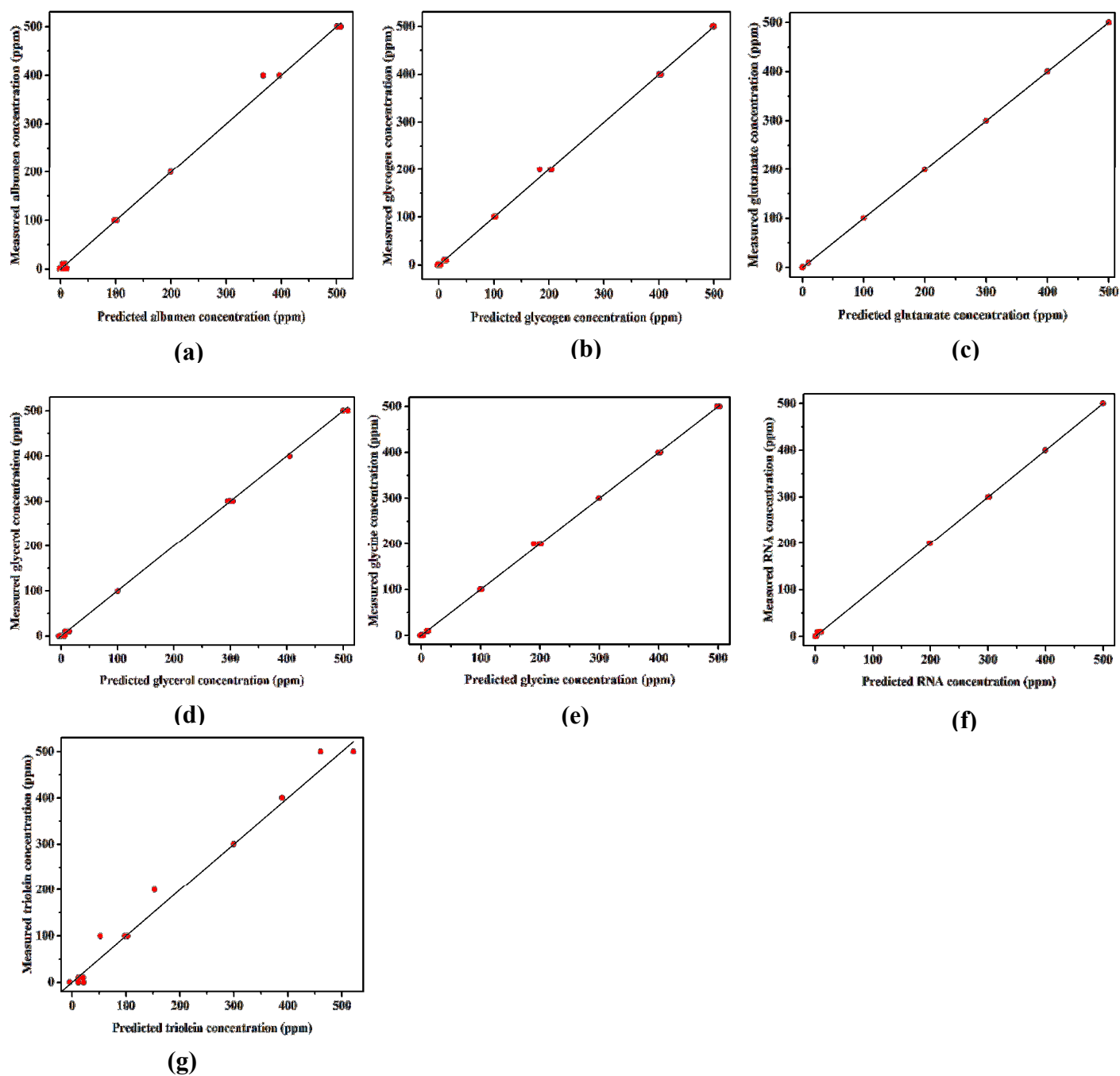


Figure 5.54 Regression plots for partial least squares measured versus predicted biochemical concentrations of the basal compounds used in the spectral model, based on the spectra profiles of whole blood samples (639 cm^{-1} , 923 cm^{-1} , 999 cm^{-1} , 1130 cm^{-1} , 1158 cm^{-1} , 1197 cm^{-1} , 1227 cm^{-1} , 1301 cm^{-1} , 1346 cm^{-1} , 1369 cm^{-1} , 1459 cm^{-1} , 1605 cm^{-1} , 1697 cm^{-1} , 1729 cm^{-1} , and 1769 cm^{-1}).

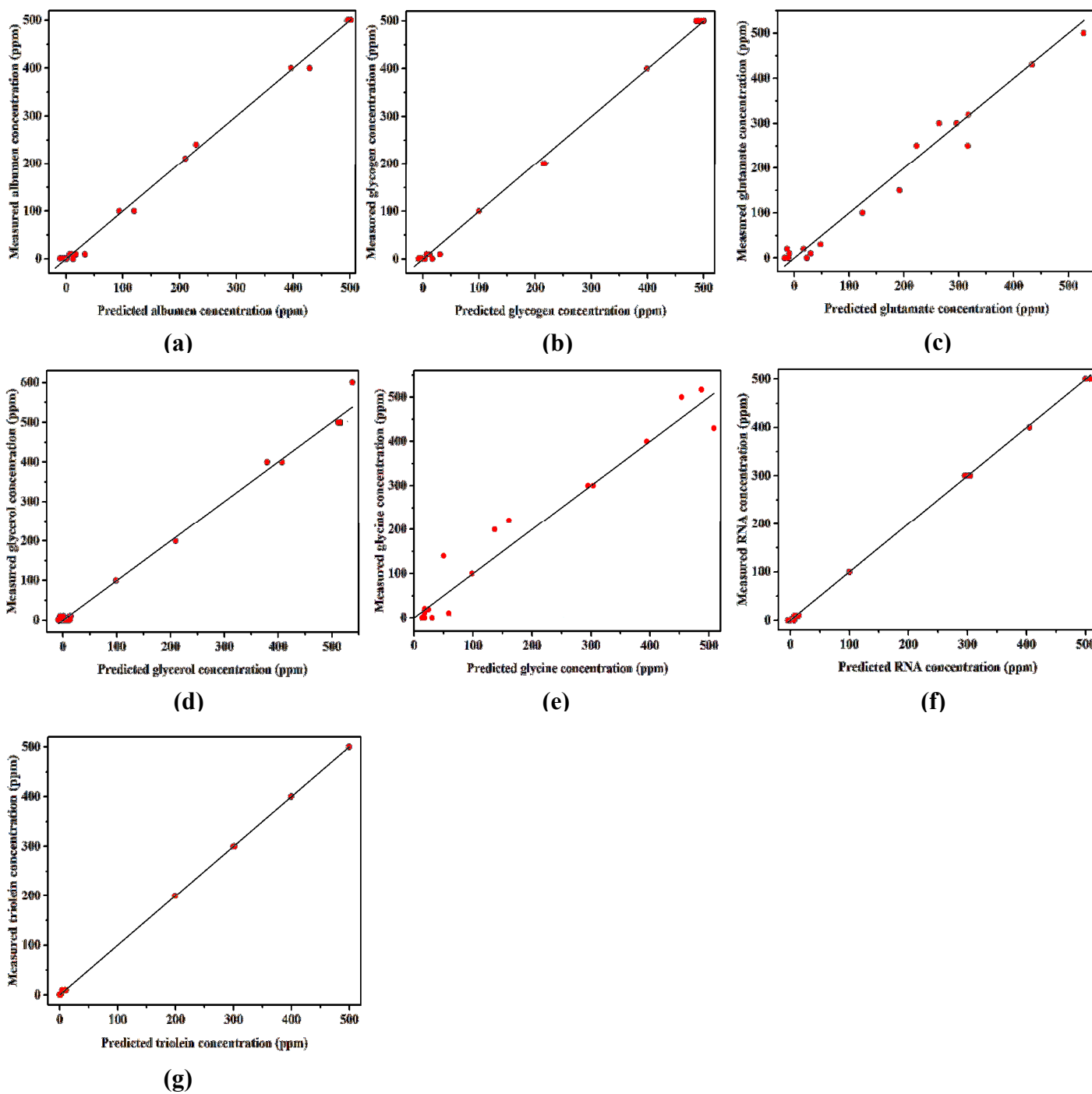


Figure 5.55 Regression plots for partial least squares measured versus predicted biochemical concentrations of the basal compounds used in the spectral model, based on the spectra profiles of saliva samples (591 cm^{-1} , 619 cm^{-1} , 652 cm^{-1} , 685 cm^{-1} , 718 cm^{-1} , 935 cm^{-1} , 1372 cm^{-1} , 1507 cm^{-1} , and 1660 cm^{-1}).

Table 5.39 Detection limits of biochemical components for Raman analysis of simulate blood fluid

| Biochemical component | Detection limits (mg / ml) | | | |
|-----------------------|----------------------------|----------------------|-------------------------|-----------------------|
| | LOD | LOQ | (<i>RMSEP</i>) | <i>R</i> ² |
| Albumen | 0.0103 | 0.031 | 0.00013 | 0.994 |
| Glycogen | 0.009 | 0.027 | 0.00001 | 0.991 |
| Glutamate | 1.254*10 ⁻⁸ | 3.8*10 ⁻⁸ | 1.825*10 ⁻⁹ | 0.997 |
| Glycerol | 0.0095 | 0.028 | 1.203*10 ⁻⁶ | 0.999 |
| RNA | 0.00098 | 0.0030 | 1.061*10 ⁻¹⁰ | 0.999 |
| Triolein | 0.0623 | 0.189 | 0.0071 | 0.902 |

Table 5.40 Detection limits of biochemical components for Raman analysis of simulate saliva fluid

| Biochemical component | Detection limits (mg / ml) | | | |
|-----------------------|----------------------------|-----------------------|-------------------------|-----------------------|
| | LOD | LOQ | (<i>RMSEP</i>) | <i>R</i> ² |
| Albumen | 0.0072 | 0.022 | 0.00172 | 0.996 |
| Glycogen | 0.0276 | 0.084 | 0.00211 | 0.984 |
| Glutamate | 1.238*10 ⁻⁸ | 3.75*10 ⁻⁸ | 1.415*10 ⁻¹⁰ | 0.932 |
| Glycerol | 0.0017 | 0.0052 | 0.00431 | 0.998 |
| Glycine | 0.00203 | 0.00615 | 0.00172 | 0.916 |
| RNA | 0.00082 | 0.0025 | 1.178*10 ⁻¹¹ | 0.997 |
| Triolein | 0.0504 | 0.153 | 0.0077 | 1 |

Table 5.41 Comparison of biochemical components concentrations in a whole blood simulate reference solution and the results obtained from PLS regression chemometric enabled Raman spectroscopy

| Biochemical Components | Concentration (mg / ml) | Measured value (\pm SD) | Deviation (%) |
|------------------------|-------------------------|----------------------------|---------------|
| Albumen | 0.4 | 0.42 \pm 0.01 | 5 |
| Glycogen | 0.1 | 0.102 \pm 0.022 | 2 |
| Glutamate | 0.001 | 0.000982 \pm 0.00013 | 0.18 |
| Glycerol | 0.01 | 0.0101 \pm 0.0016 | 1 |
| RNA | 0.002 | 0.00205 \pm 0.0001 | 2.5 |
| Triolein | 0.3 | 0.276 \pm 0.0029 | 8 |

Table 5.42 Comparison of biochemical components concentrations in a standard saliva simulate and the results obtained from PLS regression

| Biochemical Components | Concentration (mg / ml) | Measured value (\pm SD) | Deviation (%) |
|------------------------|-------------------------|----------------------------|---------------|
| Albumen | 0.4 | 0.39 \pm 0.03 | 1 |
| Glycogen | 0.1 | 0.103 \pm 0.01 | 3 |
| Glutamate | 0.001 | 0.000992 \pm 0.00024 | 0.8 |
| Glycerol | 0.01 | 0.0102 \pm 0.0024 | 2 |
| RNA | 0.002 | 0.00203 \pm 0.00011 | 1.5 |
| Triolein | 0.3 | 0.294 \pm 0.0036 | 2 |

Table 5.43 Estimated amounts of biochemical components in whole blood of normal (control) and grade 3 leukemia patients - based on the fingerprint (500-1800 cm^{-1}) and the selected (639 cm^{-1} , 923 cm^{-1} , 999 cm^{-1} , 1130 cm^{-1} , 1158 cm^{-1} , 1197 cm^{-1} , 1227 cm^{-1} , 1301 cm^{-1} , 1346 cm^{-1} , 1369 cm^{-1} , 1459 cm^{-1} , 1605 cm^{-1} , 1697 cm^{-1} , 1729 cm^{-1} , 1769 cm^{-1}) spectral regions

(a)

| 500-1800 cm^{-1} region | | Biochemical components (ppm) | | | | | |
|----------------------------------|---------|------------------------------|-----------|----------|---------|------|----------|
| Disease status | Albumen | Glycogen | Glutamate | Glycerol | Glycine | RNA | Triolein |
| Controls | 14.6 | 12.34 | 18.99 | 7.9 | 14.75 | 12.6 | 10.2 |
| Grade 3 | 25.8 | 17.93 | 14.93 | 18.8 | 25.15 | 38.9 | 11.1 |

(b)

| Based on subtle band regions | | Biochemical components (ppm) | | | | | |
|------------------------------|---------|------------------------------|-----------|----------|---------|-------|----------|
| Disease status | Albumen | Glycogen | Glutamate | Glycerol | Glycine | RNA | Triolein |
| Controls | 4.04 | 2.72 | 2.29 | 4.32 | 14.7 | 15.61 | 7.1565 |
| Grade 3 | 6.14 | 2.8 | 1.89 | 2.21 | 11.1 | 32.25 | 3.9135 |

Table 5.44 Estimated amounts of biochemical components in saliva of normal (control) and grade 3 leukemia patients - based on the fingerprint (500-1800 cm^{-1}) and the subtle bands (591 cm^{-1} , 619 cm^{-1} , 652 cm^{-1} , 685 cm^{-1} , 718 cm^{-1} , 935 cm^{-1} , 1372 cm^{-1} , 1507 cm^{-1} , and 1660 cm^{-1}) spectral regions

(a)

| 500-1800 cm^{-1} region | | Biochemical components (ppm) | | | | | |
|----------------------------------|---------|------------------------------|-----------|----------|---------|-------|----------|
| Disease status | Albumen | glycogen | glutamate | glycerol | glycine | RNA | Triolein |
| Controls | 17.3 | 23.7 | 10.84 | 25.6 | 15.4 | 4.2 | 23.19 |
| Grade 3 | 44.62 | 47.3 | 52.17 | 40.96 | 30.74 | 17.26 | 66.04 |

(b)

| Based on subtle band regions | | Biochemical components (ppm) | | | | | |
|------------------------------|---------|------------------------------|-----------|----------|---------|-------|----------|
| Disease status | Albumen | Glycogen | Glutamate | Glycerol | Glycine | RNA | Triolein |
| Controls | 11.39 | 14.90 | 1.72 | 4.769 | 5.04 | 1.069 | 1.81 |
| Grade 3 | 8.737 | 7.82 | 15.88 | 3.645 | 17.80 | 5.077 | 0.282 |

If we consider the fingerprint region (500-1800 cm^{-1}), it can be seen (Table 5.43 (a), Table 5.44 (a)) that the relative amounts of the selected biochemical components were greater in Raman spectra of leukemia patients when compared to Raman spectra of control patients, meaning that the total amounts of proteins, nucleic acids and saccharides were greater in leukemia patients. However, quantification of the biochemical components in blood Raman spectra using the selected subtle band regions, it was observed (Figure 5.43 (b)) that the relative amounts of albumen and RNA were greater in Raman spectra of leukemia patients when compared to Raman spectra of normal (control) patients whereas the relative amounts of glutamate, glycerol, glycine and triolein were greater in Raman spectra of control patients when compared to amounts in Raman spectra of leukemia patients. In the context of the selected markers, this suggests that the amount of proteins in leukemia patients were greater when compared to amounts in healthy volunteers. This can be attributed to tentative assignments corresponding to tyrosine (639 cm^{-1}), proline / valine (923 cm^{-1}) phenylalanine (999 cm^{-1}), carotenoids (1158 cm^{-1}), and amide III (1197 cm^{-1}). Similarly, greater amounts of RNA in leukemia patients implies that leukemia samples had greater amounts of nucleic acids. This is well understood given the considered number of nucleic acid markers in leukemia samples (1227 cm^{-1} , 1346 cm^{-1} , 1459 cm^{-1}) in comparison to the number of nucleic acid markers in control samples (1605 cm^{-1}). Notably, examination of Figure 5.53 (c) shows that the spectral marker at 1227 cm^{-1} pointed to presence of leukemia. The spectra markers at 1225 - 1245 cm^{-1} are associated with phosphate stretching modes that originate from the phosphodiester groups of nucleic acids, known to suggest an increase in the nucleic acids in the malignant tissues (Movasaghi *et al.*, 2007).

Visual examination of Table 5.44 (b) shows that the relative amounts of glutamate, glycine, triolein and albumen were greater in saliva samples of leukemia patients when compared to normal (control) patients, whereas the relative amounts of glycogen, glycerol, and RNA were greater in samples of control patients when compared to diseased patients. It can be concluded that biochemical alterations associated with the selected spectral markers signify that proteins and membranous lipids were greater in leukemia patients whereas nucleic acids, glycogen and non-membranous lipids were greater in control patients.

5.3.4 Multivariate statistical analysis of blood and saliva spectra for leukemia diagnostics

The SVM models are prominent for handling both linear and non-linear data. The model aims to draw decision boundaries between data points from different classes and separate them with maximum margin (Christopher *et al.*, 1998). Due to the non-linear nature (multiple type and kind of patterns) of spectra datasets in the leukemia study, the radial basis function (RBF) SVM and backpropagation neural network (BPNN) classifiers were selected for chemometric analysis. For blood spectra datasets, 15 neurons per layer, learning rate=0.01, and number of iterations = 1000 were used as variables in the BPNN for the construction of a predictive model, whereas the first ten PCs, explaining the 97 % of spectral variance, were used as variables in the RBF-SVM for the construction of a predictive model, from which the sensitivity, specificity, and overall accuracy of the method can be estimated. Figure 5.56 shows the scatter plots of RBF-SVM and BPNN diagnostic models demonstrating clustering of Raman spectra of blood samples from normal (control) and leukemia patients. The results of the ten-fold cross- validated RBF-SVM and BPNN classification are reported, in the form of a confusion matrix, in Table 5.45 and Table 5.46, respectively.

For analysis of saliva datasets, 10 neurons per layer, learning rate=0.01, and number of iterations = 1000 were used as variables in the BPNN for the construction of a predictive model. Moreover, kernel parameter = 0.4, cost=100, \approx 148 support vectors and five principal components were used for construction of RBF-SVM predictive models. Figure 5.57 shows the scatter plots of RBF-SVM and BPNN diagnostic models demonstrating clustering of Raman spectra of saliva samples from normal (control) and leukemia patients. The estimated sensitivity, specificity, and overall accuracies of the RBF-SVM and BPNN methods are summarized in Table 5.47 and Table 5.48, respectively.

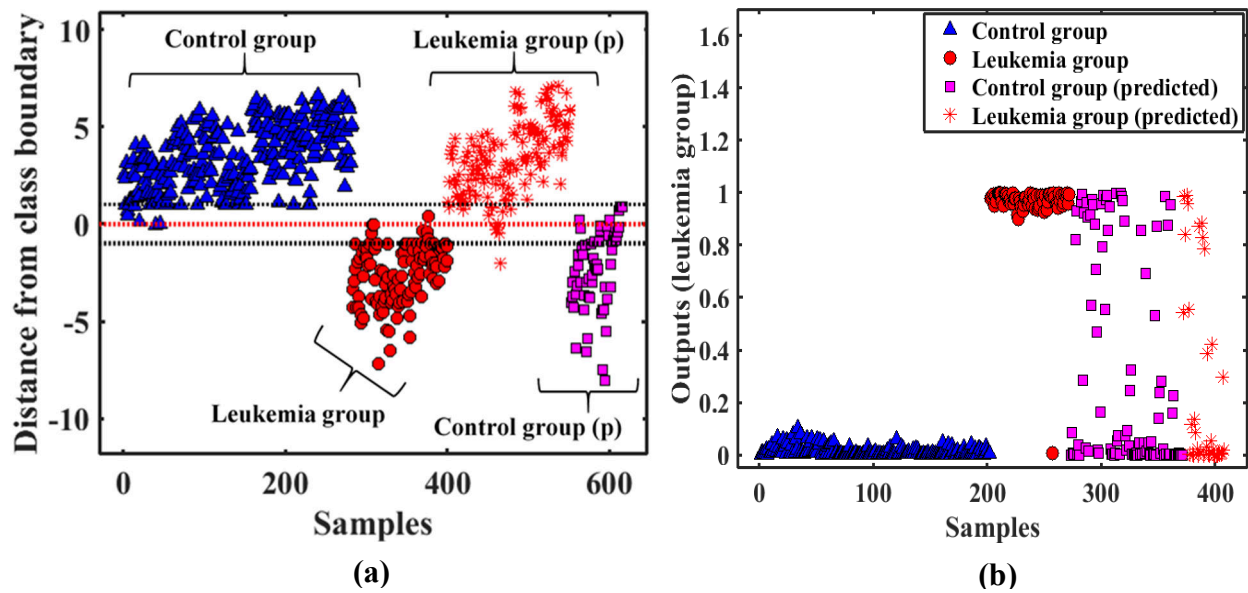


Figure 5.56 Scatter plots of (a) RBF-SVM and (b) BPNN diagnostic models demonstrating clustering of Raman spectra of blood samples from healthy volunteers and leukemia patients.

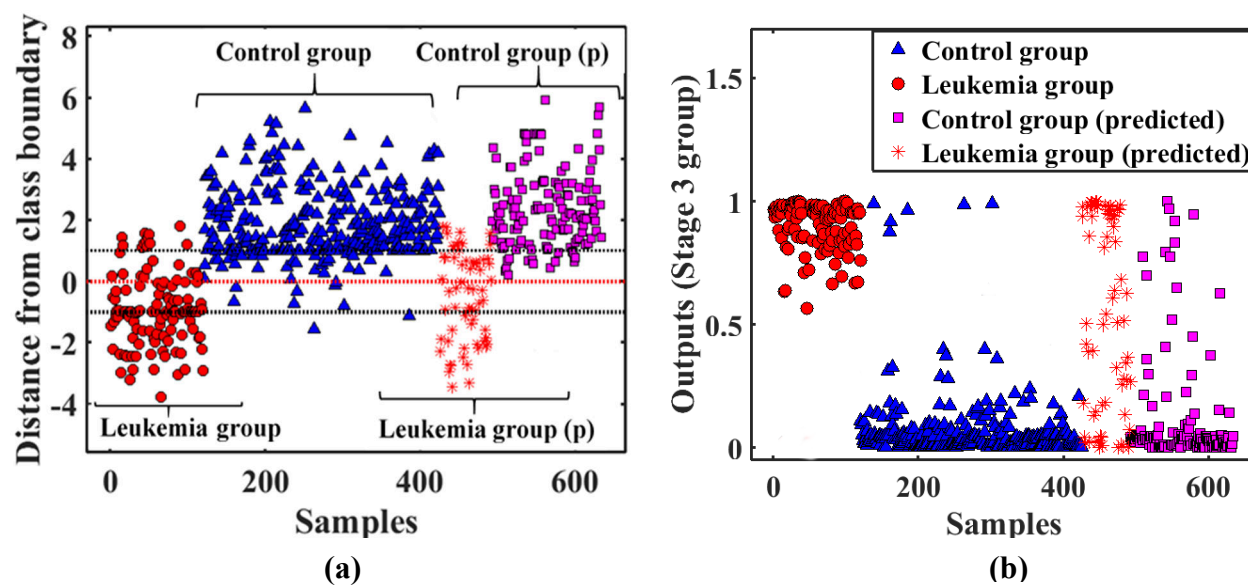


Figure 5.57 Scatter plots of (a) RBF-SVM and (b) BPNN diagnostic models demonstrating clustering of Raman spectra of saliva samples from healthy volunteers and leukemia patients.

Table 5.45 Diagnostic results of RBF-SVM predictor model on the Raman spectra of blood samples from healthy volunteers (controls) and leukemia patients

| | | Cases | | | | | | |
|----------------|-----------|---------------|----------|--------------|-------|--------------|-----------------|-----------------|
| Disease status | Diagnosis | Breast cancer | Controls | Not assigned | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
| Stage-3 | Leukemia | 57 | 7 | 0 | 64 | 94 | 89 | 97 |
| | Controls | 5 | 147 | 0 | 152 | | | |

Table 5.46 Diagnostic results of BPNN predictor model on the Raman spectra of blood samples from healthy volunteers (controls) and leukemia patients

| | | Cases | | | | | | |
|----------------|-----------|---------------|----------|--------------|-------|--------------|-----------------|-----------------|
| Disease status | Diagnosis | Breast cancer | Controls | Not assigned | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
| Stage-3 | Leukemia | 37 | 27 | 0 | 64 | 67 | 58 | 71 |
| | Controls | 45 | 107 | 0 | 152 | | | |

Table 5.47 Diagnostic results of RBF-SVM predictor model on the Raman spectra of saliva samples from healthy volunteers (controls) and leukemia patients

| | | Cases | | | | | | |
|----------------|-----------|---------------|----------|--------------|-------|--------------|-----------------|-----------------|
| Disease status | Diagnosis | Breast cancer | Controls | Not assigned | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
| Stage-3 | Leukemia | 50 | 17 | 0 | 67 | 89 | 76 | 96 |
| | Controls | 5 | 137 | 0 | 142 | | | |

Table 5.48 Diagnostic results of BPNN predictor model on the Raman spectra of saliva samples from healthy volunteers (controls) and leukemia patients

| | | Cases | | | | | | |
|----------------|-----------|---------------|----------|--------------|-------|--------------|-----------------|-----------------|
| Disease status | Diagnosis | Breast cancer | Controls | Not assigned | Total | Accuracy (%) | Sensitivity (%) | Specificity (%) |
| Stage-3 | Leukemia | 39 | 27 | 1 | 67 | 80 | 59 | 90 |
| | Controls | 14 | 128 | 0 | 142 | | | |

As observed (Table 5.45 - 5.48), the RBF-SVM model performed better than the BPNN model in diagnosing and predicting leukemia, using Raman spectra from either blood or saliva samples. This can be attributed to non-parametric nature of RBF function which strengthens its ability in handling complex data (Chen *et al.*, 2015; Sajda, 2006; Mika *et al.*, 2002). Utility of saliva spectra in RBF-SVM and BPNN diagnostic predictor models led to poor diagnostic capabilities in terms of sensitivity parameters, possibly due to inherently small scattering cross-section and the strong background fluorescence interference of Raman technique on saliva samples (Feng *et al.*, 2015), which most likely make the technique not sensitive enough for detecting the subtle biochemical changes in human saliva samples for medical diagnosis. On a positive note, application of RBF-SVM and BPNN diagnostic models on blood spectra yielded the higher diagnostic parameters, leading to a sensitivity of 89 %, a specificity of 97 %, and an overall diagnostic accuracy of 94 %. These results demonstrate that the RBF-SVM-based blood Raman spectral classification method is powerful for the diagnosis of leukemia.

6.0 Conclusions and Recommendations

A number of findings presented in this study are of great significance and have been reported here for the first time. The results obtained on data reported in this study support the idea that analysis of higher-order principal components is a novel multivariate analysis method of understanding trace biomarker alterations that point to breast cancer and leukemia progression. The implications of this finding are that with suitable cell lines representing breast and leukemia malignancy, the proposed method can be extended to study the onset stages of cancer development *in-vitro*.

When considering the first two aims of this study, we can conclude that our first aim - to identify and determine the concentrations of trace biomarkers of leukemic and breast cancer in saliva and blood using laser Raman microspectroscopy, and the second aim- to correlate the obtained biomarker levels as well as their alterations in the selected body fluids matrices to cancer presence and severity based on concentration levels of biochemical changes and the band ratios of trace spectral markers - have largely been met. Spectral analysis on Raman spectra of blood and saliva from healthy volunteers and breast cancer patients revealed that biochemical differences in healthy and diseased samples were mainly due to proteins, lipids, and nucleic acid components. Six spectral regions (subtle markers) were determined: 589 cm^{-1} , 594 cm^{-1} , 630 cm^{-1} , 1626 cm^{-1} , 1630 cm^{-1} and 1638 cm^{-1} , which can be used as new biomarkers of breast cancer. Evaluation of biochemical changes at trace peaks regions (589 , 594 , 630 , 858 , 868 , 1005 , 1160 , 1250 , 1347 , 1358 , 1626 , 1630 , and 1638 cm^{-1}) with the developed partial least-squares fitting regression model showed concentrations of proteins, nucleic acids, and lipid levels increased with breast malignancy. Moreover, these regions differentiated diseased from normal samples with acceptable levels of sensitivity using the PLS-DA algorithm. The number of correctly identified cases out of total cases led to an accuracy of 98%, 98% and 94% for grade-1, grade -2 and grade -3 cancers, respectively. The sensitivity, expressed as the number of correctly identified cancer spectra over the total number of diseased spectra was found to be 100% for grade 1 cancer, 98% for grade 2 cancer and 94% for grade 3 cancer. The specificity, expressed as the number of correctly identified healthy (control) spectra over the total number of healthy spectra was determined to be >96%, for all considered stages of cancer.

With regard to saliva analysis, analysis of lower and intermediate analysis yielded twelve spectral regions: $643\text{-}647\text{ cm}^{-1}$, $687\text{-}689\text{ cm}^{-1}$, $816\text{-}818\text{ cm}^{-1}$, $1022\text{-}1024\text{ cm}^{-1}$, $1125\text{-}1128\text{ cm}^{-1}$,

1145-1148 cm^{-1} , 1164-1166 cm^{-1} , 1427-1430 cm^{-1} , 1570-1572 cm^{-1} , 1609-1619 cm^{-1} , 1630-1657 cm^{-1} , and 1753-1756 cm^{-1} , which were regarded as trace Raman peaks for further analysis. A statistical analysis on mean and standard deviations of subtle markers showed that changes in ring breathing modes of DNA bases ($690 \pm 1.47 \text{ cm}^{-1}$), tyrosine proteins ($1159 \pm 4.24 \text{ cm}^{-1}$), amide II proteins ($1570 \pm 0.47 \text{ cm}^{-1}$), and amide I proteins ($1644 \pm 4.73 \text{ cm}^{-1}$) increased with breast cancer progression, whereas the changes in proline / tyrosine proteins (817 ± 0.40 , $1614 \pm 2.04 \text{ cm}^{-1}$) and lipids ($1754 \pm 0.23 \text{ cm}^{-1}$) decreased with breast cancer progression. Determination of biochemical components associated with the observed trace Raman peaks; using PLS regression method, suggested that amounts of glycogen decreased with progression of malignancy, whereas the amounts of proteins, nucleic acids and adipocyte levels of membranous lipids increased with malignancy. Stage wise comparison of breast cancer was performed by PLS-DA with the $k = 10$ fold cross-validation method. The developed PLS-DA algorithm achieved diagnostic sensitivities of 93%, 91%, and 91%; specificities of 96%, 93%, and 91%; and accuracies of 96%, 92%, and 91%, respectively, when differentiating normal saliva samples from grade 1 saliva samples, grade 2 saliva samples, and grade 3 saliva samples. This strengthens the view that complexity of comparing biochemical and morphological alterations amongst the diseased and normal / control samples increase with cancer progression.

Our third aim- to apply robust and hybridized machine learning techniques (higher-order PCA, ICA, MDS, PLS-DA, and kernel density estimators) in the extraction and multivariate exploratory analysis and interpretation of the biomarkers embedded in the measured spectra – has also been largely met. Analysis of blood spectra datasets in healthy volunteers and breast cancer patients using ICA revealed that the sum of eigenvalues (in percentage) for the number of retained eigenvalues decreased with stage of cancer progression. This suggests there were additional spectral regions in respective datasets that could be characteristically considered as noise and therefore could not be useful for breast cancer diagnosis (Crow *et al.*, 2005). Analysis of ICA loading vectors showed aromatic acids proteins were a major factor in clustering of both healthy and diseased samples, meaning blood protein degradation is a major factor in breast cancer progression. ICA followed by PLS-DA performed better than PLS-DA alone in revealing trace spectral markers that were responsible for discriminating control from diseased samples, potentially due to characteristic property of ICA in ensuring statistical independence of markers, hence, minimal overlapping of biochemical information (Yao *et al.*, 2012). The implications of this finding reinforce the capability of ICA in producing producing basis vectors that are

statistically independent and not just linearly decorrelated as it happens with PCA. Further, ICA-MDS followed by PLS-DA marginally yielded a better diagnosis of late (grade 3) malignancy at sensitivity of 97% when compared to sensitivity at 96% achieved by ICA followed by PLS-DA. The better performance of MDS can be attributed to its strength in mapping all pairwise distances between data points into small dimensional Euclidean domains (Aflalo *et al.*, 2013), while preserving the intrinsic information of pairwise dissimilarities between objects (Liu *et al.*, 2019).

Similarly, a combination of machine learning techniques of ICA, MDS, PLS-DA and kernel density estimators were used for analysis of saliva spectra. The analysis was performed on spectral matrices measured in 643-647 cm^{-1} , 687-689 cm^{-1} , 816-818 cm^{-1} , 1022-1024 cm^{-1} , 1125-1128 cm^{-1} , 1145-1148 cm^{-1} , 1164-1166 cm^{-1} , 1427-1430 cm^{-1} , 1570-1572 cm^{-1} , 1609-1619 cm^{-1} , 1630-1657 cm^{-1} , and 1753-1756 cm^{-1} regions. A greater number of eigenvalues were needed for optimal ICA analysis on saliva spectra when compared to blood spectra, meaning that saliva datasets were transformed into many directions of new feature spaces (and therefore magnitudes), suggesting that spectral biochemical components of saliva were complex by nature in comparison to components in blood samples. ICA followed by PLS-DA yielded diagnostic sensitivities of 89%, 95% and 92%, with a specificity of 95%, 95% and 92%, for grade 1, grade 2 and grade 3 breast cancers, respectively. Different from diagnostic results of ICA followed by PLS-DA, utility of ICA followed by MDS and kernel density estimators yielded diagnostic sensitivity of 96%, 98% and 94%; specificity of 99%, 98% and 95%, for the breast cancer patients and the healthy volunteers, respectively. These results confirm the outstanding diagnostic accuracy of the ICA-MDS-kernel density estimators-based diagnostic algorithm for breast cancer detection.

When considering the last two aims of this study, we can conclude that our fourth aim - to develop conceptual diagnostic models to detect and characterize breast and leukemia cancers in their various stages based based on support vector machine (SVM) and artificial neural networks (ANN), and the fifth aim- to test the developed diagnostic models for proof of concept, to detect and predict the status of breast and leukemia cancers in clinical liquid biopsies samples- have been fully met. In general, we observed that overall diagnostic accuracies and sensitivities decreased with malignancy, meaning underlying complexity of biochemical alteration in healthy and diseased samples increased with malignancy. Moreover, analysis of blood and saliva spectra for breast cancer diagnostic showed that a greater number of principal components and support vectors were needed for advanced stages of cancer (grade 3) in comparison to number of principal components and support vectors needed for analysis of early malignancy (grade 1, grade 2). This

suggests that late malignancy matrices greatly suffered from problems of high dimensionality and collinearity, thus, consideration of higher number of principal components was necessary to account for greater amount of variance in respective datasets (Björklund, 2019). To that effect, a relatively greater number of support vectors were needed to optimally define a hyperplane for maximizing margins between the two classes (controls versus the diseased scores) (Martins *et al.*, 2009).

Analysis of blood spectra showed that RBF kernel function model performed better than linear kernel function model in diagnosing late (grade 3) malignancy. This finding imply that the linear separable characteristic nature of spectral datasets decreased with malignancy, meaning a nonparametric method that can handle more complex data relationships. Further, comparison of SVM and BPNN predictor models diagnostic performance showed that BPNN outperformed SVM in predicting diseased samples. This could be due to better parameter selection or the diverse and non-linear nature of the data set, or both.

In agreement with analysis of blood spectra, enhanced optimal performance of RBF-SVM and BPNN diagnostic models was evident in analysis of saliva spectra. Analysis of saliva spectra showed that RBF-SVM performed better than linear-SVM during model training and prediction, in terms of diagnostic accuracy and sensitivity, proving it can be a useful machine learning technique for diagnosis of breast cancer. However, both RBF-SVM and linear-SVM predictive models yielded poor performance in saliva datasets, with the best diagnostic sensitivity being 78%. This was most likely due to the low number of patient samples or / and the complexity of the model. In particular, the increase in the size or the number of parameters in the machine learning model could have contributed to overfitting that led to poor performance. In contrast, BPNN training and predictive models performed better than the linear-SVM and RBF-SVM training and predictive models.

With regard to leukemia, salivary and blood nucleic acids, proteins and lipids can be chiefly regarded as biomarkers pointing to presence of leukemia. Quantification of corresponding subtle biochemical components in blood spectra by PLS regression model showed that the amount of proteins and nucleic acids in leukemia patients were greater in diseased patients when compared to amounts in healthy volunteers. A similar analysis on trace bands of saliva spectra demonstrated that proteins and membranous lipids were greater in leukemia patients whereas nucleic acids, glycogen and non-membranous lipids were greater in control patients.

RBF-SVM model performed better than the BPNN model in diagnosing and predicting leukemia, upon analysing Raman spectra of blood and saliva samples. We can attribute RBF-SVM better performance to non-parametric nature of RBF function that strengthen its ability in handling complex data (Chen *et al.*, 2015; Sajda, 2006; Mika *et al.*, 2002). Utility of saliva spectra in RBF-SVM and BPNN diagnostic models yielded poor diagnostic capabilities in terms of sensitivity parameters. This could be due to inherently small scattering cross-section and the strong background fluorescence interference of Raman technique on saliva samples (Feng *et al.*, 2015), which most likely made the Raman technique not sensitive enough for detecting the subtle biochemical changes in human saliva samples (Feng *et al.*, 2015). These findings strengthen the view that the RBF-SVM-based blood Raman spectral classification method can be a powerful technique for the diagnosis of leukemia.

This study is preliminary work as the complex structure of saliva and blood requires many other future investigations to find more information about changes that would occur in saliva and blood components during carcinogenesis. Much additional research still has to be done in order to elucidate the conformational properties of the nucleic acids, proteins and lipids in these samples. Importantly, knowing that bands corresponding to nucleic acids, proteins, and phospholipids played a key role in breast cancer and leukemia progression in our study, is not enough. Infact, the pairwise comparison of mean intensity (peak intensity ratios) revealed there were changes in concentration of biomolecules during cancer progression. However, from a histochemical perspective, our results could not determine exactly what molecules e.g., μ RNA, μ DNA, μ protein biomarkers (e.g., CA15-3, c-erB2, HSP90A) were responsible for the biochemical differences. Other techniques such as liquid chromatography, mass spectrometry, and enzyme-linked immunosorbent assay (ELISA) could be used to acquire complementary information for Raman microspectroscopy analysis.

While we achieved high overall accuracies through SVM and ANN models, low diagnostic accuracies, particularly for saliva datasets have remained as a challenge, which might be due to the high biochemical similarities within normal and diseased samples, or, perhaps, the variations in Raman spectra of saliva associated with multiple donors may have been significantly larger than variations found for blood samples (Sikirzhytski *et al.*, 2011). The limited number of samples from willing volunteers per cancers in consideration not only restricted us in sample sizes, but also made it harder to achieve 100% diagnostic capability; both in accuracy and sensitivity. Besides the quantity of spectra, strong background fluorescence interference of Raman spectra on saliva

samples appeared to be another issue that might have introduced artifacts in the feature extraction stage. Even though we compared prediction models under the same conditions, this could be far from a real-world scenario, considering the potential effects of slight changes in experimental conditions of Raman spectroscopy, and the limitations of inclusion and exclusion criteria governing recruitment of research participants. We believe this research can be improved further when future studies use a larger data set with more features, and validating the findings using large independent cohorts of patients before translation of the studies to a wide range of clinical applications. Moreover, better results can be achieved if datasets comprising both spatial and spectral information are included in the study. This will aid in analysis of morphological changes during progression of malignancy.

References

- Abdelrahman, A. H., Ahmed, M. A., Medani, I. M. K., and Izldeen, M. (2014). Interpretation of Raman Effect on the Basis of Quantum Theory. *Academic Research International*, 5(2): 107–112.
- Aflalo, Y., and Kimmel, R. (2013). Spectral multidimensional scaling. *Proceedings of the National Academy of Sciences of the United States of America*, 110(45): 18052–18057.
- Agha-Hosseini, F., Mirzaii-Dizgah, I., and Rahimi, A. (2009). Correlation of serum and salivary CA15-3 levels in patients with breast cancer. *Medicina, Oral, Patologia, Oral Y Cirugia, Bucal*, 10(6): 521–524.
- Allegra, A., Alonci, A., Campo, S., Penna, G., Petrunaro, A., Gerace, D., and Musolino, C. (2012). Circulating microRNAs : New biomarkers in diagnosis , prognosis and treatment of cancer (Review). *International Journal of Oncology*, 41: 1897–1912.
- Allegrini, F., and Olivieri, A. C. (2014). IUPAC-consistent approach to the limit of detection in partial least-squares calibration. *Analytical Chemistry*, 86(15): 7858–7866.
- American Cancer Society. (2015). Global Cancer Facts and Figures 3rd Edition. In *Atlanta: American Cancer Society*. Atlanta, Georgia, 1-26.
- American Cancer Society. (2017). Cancer facts and figures 2017. In *American Cancer Society*. Atlanta, 1-30.
- American Cancer Society. (2011). *Global cancer facts and figures*. 2nd Edition, Atlanta, 1–60,
- Antabe, R., Kansanga, M., Sano, Y., Kyeremeh, E., and Galaa, Y. (2020). Utilization of breast cancer screening in Kenya: What are the determinants? *BMC Health Services Research*, 20(1): 1–9.
- Archibald, R., and Fann, G. (2007). Feature selection and classification of hyperspectral images with support vector machines. *IEEE Geoscience And Remote Sensing Letters*, 4(4): 674–677.
- Ashok, L., Gp, S., and Hema, G. (2010). Estimation of salivary amylase and total proteins in leukemia patients and its correlation with clinical feature and radiographic finding. *Indian Journal in Dental Research*, 21(4): 486–490.
- Atkins, C. G., Buckley, K., Blades, M. W., and Turner, R. F. B. (2017). Raman spectroscopy of blood and blood components. *Applied Spectroscopy*, 71(5): 767–793.

- Avram, L., Stefancu, A., and Crisan, D. (2020). Recent advances in surface - enhanced Raman spectroscopy based liquid biopsy for colorectal cancer (Review). *Experimental and Therapeutic Medicine*, 213: 1-7.
- Awad, M., and Khanna, R. (2015). Efficient learning machines: Theories, concepts, and applications for engineers and system designers. *Chemometrics*, 3(2): 1–248.
- Babrah, J., Mccarthy, K. P., Lush, R., Rye, A. D., Bessant, C., and Stone, N. (2007). FT-Infrared spectroscopic studies of lymphoma, lymphoid and myeloid leukemic cell lines. *SPIE-OSA Biomedical Optics: Diagnostic Optical Spectroscopy in Biomedicine IV*, 6628: 1–7.
- Baker, M. J., Hussain, S. R., Lovergne, L., Untereiner, V., Hughes, C., Lukaszewski, R. A., and Sockalingum, G. D. (2016). Developing and understanding biofluid vibrational spectroscopy: A critical review. *Chemical Society Reviews*, 45(7): 1803-1818.
- Bauer, D. E., Hatzivassiliou, G., Zhao, F., Andreadis, C., and Thompson, C. B. (2005). ATP citrate lyase is an important component of cell growth and transformation. *Oncogene*, 24(41): 6314–6322.
- Bearman, G., and Levenson, R. (2003). Biological imaging spectroscopy. In *Biomedical Photonics: Handbook*, 3rd edition, New York, 342-366.
- Beata, B., Jacek, M., Radzislav, K., Elena B., and Thomas D., H. A. (2012). Raman spectroscopy and imaging : applications in human breast cancer diagnosis. *Analyst*, 137: 3773–3780.
- Belousov, A. I., Verzakov, S. A., and Frese, J. Von. (2002). A flexible classification approach with optimal generalisation performance : support vector machines. *Chemometrics and Intelligent Laboratory Systems*, 64: 15–25.
- Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). *Biochemistry*, 5th edition. New York: W.H. Freeman, 135-139.
- Bilal, M., Tabassum, S., Saleem, M., Mahmood, H., Sarwar, U., and Ullah K. E. (2017). Optical Screening of Female Breast Cancer from Whole Blood using Raman spectroscopy. *Applied Spectroscopy*, 71(5): 1-10.
- Bird, B., Miljkovic, M., Romeo, M. J., Smith, J., Stone, N., George, M. W., and Diem M. (2008). Infrared micro-spectral imaging : distinction of tissue types in axillary lymph node histology. *BMC Clinical Pathology*, 8(8): 1-14.
- Bisgin, H., Bera, T., Ding, H., Semey, H. G., Wu, L., Liu, Z., and Xu, J. (2018). Comparing SVM and ANN based Machine Learning Methods for Species Identification of Food Contaminating Beetles. *Scientific Reports*, 8 (6532): 1–12.

- Bishop, C. M., Eng, C. M. B. F. R., and Jordan, M. (2006). *Pattern recognition and machine learning*. Singapore: Springer Science and Business Media LLC, 650-678.
- Björklund, M. (2019). Be careful with your principal components. *Evolution*, 73(10): 2151–2158.
- Boiret, M., Rutledge, D. N., Gorretta, N., Ginot, Y. M., Roger, J. M., and Boiret, J. M. R. (2014). Application of independent component analysis on raman images of a pharmaceutical drug product: pure spectra determination and spatial distribution of constituents. *Journal of Pharmaceutical and Biomedical Analysis*, 90: 78–84.
- Bouzalmat, A., and Kharroubi, J. (2014). Comparative Study of PCA , ICA , LDA using SVM Classifier. *Emerging Technologies in Web Intelligence*, 6(1): 64–68.
- Brozek-Pluska, B., Musial, J., Kordek, R., Bailo, E., Dieing, T., and Abramczyk, H. (2012). Raman spectroscopy and imaging: Applications in human breast cancer diagnosis. *Analyst*, 137(16): 3773–3780.
- Brozoski, D. T., and Santos, C. F. (2017). Human DNA extraction from whole saliva that was fresh or stored for 3 , 6 or 12 months using five different protocols. *Journal of Applied Oral Science*, 25(2): 147–158.
- Brunner, D., Frank, J., Appl, H., Schoffl, H., Pfaller, W., and Gstraunthaler, G. (2009). Serum-free cell culture : The serum-free media interactive online database. *Altex*, 27 (10): 53–62.
- Bryne, H., Knief, P., Keating, M., and Bonnier, F. (2015). Spectral pre and post processing for Infrared and Raman spectroscopy of biological tissues and cells. *Chemical Science*, 0(3): 1–12.
- Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2: 121-167.
- Byrne, H. J., Bonnier, F., McIntyre, J., and Rajan, D. (2020). Quantitative analysis of human blood serum using vibrational spectroscopy. *Clinical Spectroscopy*, 2: 1–15.
- Byrne, H., Kerr, L., and Hennelly, B. M. (2015). Optimal choice of sample substrate and laser wavelength for Raman spectroscopic analysis of biological specimen. *Analytical Methods*, 7: 1–14.
- Calado, G., and Lyng, F. M. (2019). Raman spectroscopic analysis of saliva for the diagnosis of oral cancer : A systematic review. *Translational Biophotonics*, 5(2): 1–10.
- Camps-valls, G., and Bruzzone, L. (2005). *Kernel-Based Methods for Hyperspectral Image Classification*. 43(6): 1351–1362.

- Cervo, S., Mansutti, E., Mistro, G. D., Colombatti, A., Steffan, A., Sergo, V., and Bonifacio, A. (2015). SERS analysis of serum for detection of early and locally advanced breast cancer. *Analytical and Bioanalytical Chemistry*, 407(24): 7503-7509.
- Chandra, A., Talari, S., Movasaghi, Z., and Rehman, S. (2015). Raman Spectroscopy of Biological Tissues. *Applied Spectroscopy Reviews*, 50: 46–111.
- Chen, D., Song, N., Ni, R., Zhao, J., Hu, J., Lu, Q., and Li, Q. (2014). Saliva as a sampling source for the detection of leukemic fusion transcripts. *Translational Medicine*, 1(321–327): 1–5.
- Chen, H., Lin, Z., and Tan, C. (2015). Cancer Discrimination Using Fourier Transform Near-Infrared Spectroscopy with Chemometric Models. *Chemistry*, 3: 1–9.
- Chiappin, S., Antonelli, G., Gatti, R., and Palo, E. F. De. (2007). Saliva specimen: A new laboratory tool for diagnostic and basic investigation. *Clinica Chimica Acta*, 383: 30–40.
- Chowdary, M. V. P., Kumar, K. K., Kurien, J., Mathew, S., and Krishna, C. M. (2006). Discrimination of Normal, Benign, and Malignant Breast Tissues by Raman Spectroscopy. *Biopolymers*, 83: 556–569.
- Christodoulides, N., Mohanty, S., Miller, C. S., Langub, M. C., Floriano, P. N., Dharshan, P., and McDevitt, J. T. (2005). Application of microchip assay system for the measurement of C-reactive protein in human saliva. *Lab on a Chip*, 5(3): 261–269.
- Chung, S.H., Park, C.S., and Park, K.S. (2005). Application of independent component analysis (ICA) method to the Raman spectra processing. *Optical Diagnostics and Sensing*, 5702: 168-172.
- Ci, Y. U. N. X., Gao, T. I. Y. U., Feng, J. U. N., and Guo, Z. Q. (1999). Fourier Transform Infrared Spectroscopic Characterization of Human Breast Tissue: Implications for Breast Cancer. *Applied Spectroscopy*, 53(3): 312–315.
- Coomans, D., Massart, D. L., and Broeckeaert, I. (1981). Potential methods in pattern recognition. *Analytica Chimica Acta*, 133: 215–224.
- Cordells, C. B. Y. (2012). PCA: The Basic Building Block of Chemometrics. 1-46.
- Corsetti, S., Rabl, T., McGloin, D., and Ghulam, N. (2018). Raman spectroscopy for accurately characterizing biomolecular changes in androgen-independent prostate cancer cells. *Biophotonics*, 11: 1–8.
- Crow, P., Barrass, B., Kendall, C., Wright, M., Persad, R., and Stone, N. (2005). The use of Raman spectroscopy to differentiate between different prostatic adenocarcinoma cell lines. *British Journal of Cancer*, 92(12): 2166–2170.

- Crow, P., Stone, N., Kendall, C. A., Uff, J. S., Farmer, J. A. M., Barr, H., and Wright, M. P. J. (2003). The use of Raman spectroscopy to identify and grade prostatic adenocarcinoma in vitro. *British Journal of Cancer*, 89: 106–108.
- Dattalo, P. V. (2014). A demonstration of canonical correlation analysis with orthogonal rotation to facilitate interpretation. *VCU Scholars Compass, Social Work Publications*, 2(3): 1–34.
- Dehghan, F., and Giti, M. (2008). Automatic Detection of Clustered Microcalcifications in Digital Mammograms : Study on Applying Adaboost with SVM-based Component Classifiers. *30th Annual International IEEE EMBS Conference*, 4789–4792.
- Delen, D., Walker, G., and Kadam, A. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, 34(2): 113–127.
- Desimoni, E., and Brunetti, B. (2015). About Estimating the Limit of Detection by the Signal to Noise Approach. *Pharmaceutica Analytica Acta*, 6(4): 1–4.
- Desroches, J., Jermyn, M., Mok, K., Lemieux-leduc, C., Mercier, J., St-arnaud, K., and Leblond, F. (2015). Characterization of a Raman spectroscopy probe system for intraoperative brain tissue classification. *Biomedical Optics Express*, 6(7): 613–623.
- Devi, G., Devi, T. S. R., and Gunasekaran, S. (2010). FTIR spectroscopic study on benign and cancerous human breast tissues-a run chart analysis. *International Journal of Pharmaceutical Sciences Review and Research*, 2(2): 73–77.
- Duffy, M. J. (2006). Serum tumor markers in breast cancer : Are they of clinical value ? *Clinical Chemistry*, 52(3): 345–351.
- Dyson, R., Marcolli, C., Juan, A. De, Rault, M., and Maeder, M. (2004). Spectroscopic imaging and chemometrics : a powerful combination for global and local sample analysis. *Trend*, 23(1): 70–79.
- Emekli-Altufran, E., Demir, G., Kasikci, E., Tunali-Akbay, T., Pisiriciler, R., Caliskan, E., and Yarat, A. (2008). Altered biochemical parameters in the saliva of patients with breast cancer. *The Tohoku Journal of Express Medicine*, 214: 89–96.
- Erukhimovitch, V., Talyshinsky, M., Souprun, Y., and Huleihel, M. (2006). FTIR spectroscopy examination of leukemia patients plasma. *Vibrational Spectroscopy*, 40(1): 40–46.
- Farnaud, S. J. C., Kosti, O., Getting, S. J., and Renshaw, D. (2010). Saliva : Physiology and diagnostic potential in health and disease. *The ScientificWorld Journal*, 10: 434–456.
- Feng, S., Huang, S., Lin, D., Chen, G., Xu, Y., Li, Y., and Zeng, H. (2015). Surface-enhanced Raman spectroscopy of saliva proteins for the noninvasive differentiation of benign and

- malignant breast tumors. *International Journal of Nan*, 10: 537–547.
- Feng, Shangyuan, Lin, D., Lin, J., Huang, Z., Chen, G., Li, Y., and Huang, S. (2014). Saliva analysis combining membrane protein purification with surface-enhanced Raman spectroscopy for nasopharyngeal cancer detection. *Applied Physics Letters*, 104(073702): 3–8.
- Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., and Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *International Journal of Cancer*, 127(12): 2893–2917.
- Ferreira, I. C. C., Aguiar, M. G., Silva, A. T. F., Santos, L. D., Santos, D. W., Goulart, L. R., and Maia, Y. C. P. (2020). Attenuated Total Reflection-Fourier Transform Infrared Spectroscopy Analysis (AFTIR) of Saliva for Breast Cancer Diagnosis. *Journal of Oncology*, 4(2): 1-9.
- Forina, M., Armanino, C., Leardi, R., and Drava, G. (1991). A class-modelling technique based on potential functions. *Journal of Chemometrics*, 5: 435–453.
- Fotakis, C., Anglos, D., Zafirooulos, V., and Georgiou, S. V. T. (2007). *Lasers in the preservation of cultural heritage: principles and applications*. New York: CRC Press, 1-20.
- Fraumeni, J. F. (2011). Molecular epidemiology: principles and practices. Foreword. In *IARC scientific publications*. 2: 1-4.
- Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. and Torre, A. J. (2018). Global cancer statistics 2018 : GLOBOCAN Estimates of incidence and mortality worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 68: 394–424.
- Freshney, R. I. (2006). Basic Principles of Cell Culture. In G. Vunjak-Novakovic and R. I. Freshney (Eds.), *Culture of Cells for Tissue Engineering*, John Wiley & Sons, Ltd, 3-22.
- Frost, T. (2016). Quantitative analysis. In *Encyclopedia of Spectroscopy and Spectrometry*. 3rd Edition, Elsevier Ltd, 34-36.
- Gahan, P. B. (2010). Circulating nucleic acids in plasma and serum : diagnosis and prognosis in cancer. *EPMA*, 1: 503–512.
- Gautam, R., Vanga, S., Ariese F., and Umapathy S. (2015). Review of multidimensional data processing approaches for Raman and infrared spectroscopy. *EPJ Techniques and Instrumentation*, 2(8): 1-38.
- Geladi, P., and Da°bakk, E. (2016). Computational methods and chemometrics in near infrared spectroscopy. *Encyclopedia of Spectroscopy and Spectrometry*, 1(1): 350–355.
- Gelder, J. De, Gussem, K. De, Vandenabeele, P., and Moens, L. (2007). Reference database of

- Raman spectra of biological molecules. *Raman Spectroscopy*, 38: 1133–1147.
- Gmbh, Q. (2010). *QLAamp*® RNA Blood Mini Handbook For total RNA purification from human whole blood Sample and Assay Technologies, 1-13.
- Gonchukov, S., Sukhinina, A., Bakhmutov, D., and Minaeva, S. (2012). Raman spectroscopy of saliva as a perspective method for periodontitis diagnostics. *Laser Physics Letters*, 9(1): 73–77.
- Gontijo, L. C., Guimarães, E., Mitsutake, H., De Santana, F. B., Santos, D. Q., and Neto, W. B. (2014). Development and validation of PLS models for quantification of biodiesels content from waste frying oil in diesel by HATR-MIR. *Revista Virtual de Química*, 6(5): 1517–1528.
- Gonzálezsolís, J. L., Aguiñagaserrano, B. I., Martínezespínosa, J. C., and Ocegüeravillanueva, A. (2011). Stage determination of breast cancer biopsy using Raman spectroscopy and multivariate analysis. *American Institute of Physics*, 1364: 33–40.
- González-Solís, J. L., Luévano-Colmenero, G. H., and Vargas-Mancilla, J. (2013). Surface enhanced Raman spectroscopy in breast cancer cells. *Laser Therapy*, 22(1): 37-42.
- Haka, A. S., Shafer-peltier, K. E., Fitzmaurice, M., Crowe, J., Dasari, R. R., and Feld, M. S. (2005). Diagnosing breast cancer by using Raman spectroscopy. *PNAS*, 102(35): 12371–12376.
- Han, B., Du, Y., Fu, T., Fan, Z., Xu, S., Hu, C., and Xu, W. (2017). 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*, 71(2): 300–307.
- Happillon, T., Untereiner, V., Beljebbar, A., Gobinet, C., Daliphard, S., Cornillet-lefebvre, P., and Manfait, M. (2015). Diagnosis approach of chronic lymphocytic leukemia on unstained blood smears using Raman microspectroscopy and supervised classification. *Analyst*, 3(2): 1–8.
- Hassoun, M., Kose, N., Kiselev, R., Kirchberger-Tolstik, T., Schie, I.W., Krafft, C., and Popp, J. (2018). Quantitation of acute monocytic leukemia cells spiked in control monocytes using surface-enhanced Raman spectroscopy. *Analytical Methods*, 1(3): 1-8.
- Hernández-arteaga, A., Jesús, J. De, Nava, Z., Kolosovas-machuca, E. S., and Jesús, J. (2017). Diagnosis of breast cancer by analysis of sialic acid concentrations in human saliva by surface-enhanced Raman spectroscopy of silver nanoparticles. *Nano Research*, 2(3): 1–9.
- Hertzmann, A., and Fleet, D. (2012). Machine Learning and Data Mining. online: <https://www.dgp.toronto.edu/~hertzman/411notes.pdf> (accessed on 2 / 4 / 2020), 1–21.
- Hong, M., and He, G. (2017). Revision to the WHO classification of acute myeloid leukemia. *Journal of Translational Internal Medicine*, 5(2): 69–71.

- Høy, M., and Segtnan, V. (2012). MATLABb tutorial for spectroscopists. Part 8: partial least squares regression. *NIR News*, 23(3): 15–17.
- Hsu, C., Chang, C., and Lin, C. (2016). *A Practical Guide to Support Vector Classification*. 1–16.
- Hu, S., Wang, J., Meijer, J., Jeong, S., Xie, Y., Yu, T., and Wong, D. T. (2007). Salivary proteomic and genomic biomarkers for primary Sjögren’s syndrome. *Arthritis and Rheumatism*, 56(11): 3588–3600.
- Hu, Y., Jiang, T., Shen, A., Li, W., Wang, X., and Hu, J. (2007). A background elimination method based on wavelet transform for Raman spectra. *Chemometrics and Intelligent Laboratory Systems*, 85(1): 94-101.
- Huang Z., McWilliams A., Lui H., McLean D. I., Lam S., and Zang H. (2003). Near-Infrared Raman spectroscopy for optical diagnosis of lung cancer. *International Journal of Cancer*, 107: 1047–1052.
- Hughes, S. R., and Chapleau, R. R. (2019). Comparing DNA quantity and quality using saliva collection following food and beverage consumption. *BMC Research Notes*, 12(1): 165–171.
- Hyvärinen, A., and Oja, E. (2000). Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(4–5): 411–430.
- Id, R. S., Id, S. H., Graham, D. G., Wellman, R., Khan, S., Thrumurthy, S., and Lovat, L. B. (2020). An optimised saliva collection method to produce high-yield , high-quality RNA for translational research. *Planetary and Space Science*, 1–16.
- Israelsen, N. D., Hanson, C., and Vargis, E. (2015). Nanoparticle Properties and Synthesis Effects on Surface-Enhanced Raman Scattering Enhancement Factor: An Introduction. *Scientific World*, 2015: 1-3.
- Jafari, M., and Ansari-pour, N. (2019). Why , When and How to Adjust Your P Values ? *Cell*, 20(4): 604–607.
- Jolliffe, I. T. (2002). *Principal Component Analysis, Second Edition*. NewYork: Springer - Verlag. 112-118, 201-209.
- Jr, L. A. D., and Bardelli, A. (2014). Liquid biopsies : Genotyping circulating tumor DNA. *Clinical Oncology*, 32(6): 579–587.
- Jr, L. S., Leite, K. R. M., Silveira, F. L., Srougi, M., Pacheco, Zangaro, R. A., and Pasqualucci, C. A. (2014). Discrimination of prostate carcinoma from benign prostate tissue fragments in vitro by estimating the gross biochemical alterations through Raman spectroscopy. *Lasers in Medical Science*. 29(4): 1469-1477.

- Jurysta, C., Bulur, N., Oguzhan, B., Satman, I., Yilmaz, T. M., Malaisse, W. J., and Sener, A. (2009). Salivary Glucose Concentration and Excretion in Normal and Diabetic Subjects. *Biomedicine and Biotechnology*, 2(2): 1–6.
- Kaczor-Urbanowicz, K., Carreras-Presa, C. M., Kaczor, T., Tu, M., Wei, F., Farcia-Gordoy, F., and Wong, D. (2017). Emerging technologies for salivaomics in cancer detection diagnostics of cancer. *Cellular and Molecular Medicine*, 21(4): 640–647.
- Kasiulevičius, V., Šapoka, V., and Filipavičiūtė, R. (2006). Sample size calculation in epidemiological studies. *Gerontologija*, 7(4): 225–231.
- Kast, R. E., Tucker, S. C., Killian, K., Trexler, M., Kenneth, V., and Auner, G. W. (2014). Emerging technology : applications of Raman spectroscopy for prostate cancer. *Cancer and Metastasis Reviews*, 33(673): 1–21.
- Katzilakis, N., Stiakaki, E., Papadakis, A., Dimitriou, H., Stathopoulos, E., Markaki, E., and Kalmanti, M. (2004). Spectral characteristics of acute lymphoblastic leukemia in childhood. *Leukemia Research*, 28: 1159–1164.
- Kazarian, A., Blyuss, O., Metodieva, G., Gentry-maharaj, A., Ryan, A., Kiseleva, E. M., and Timms, J. F. (2017). Testing breast cancer serum biomarkers for early detection and prognosis in pre-diagnosis samples. *British Journal of Cancer*, 116(4): 501–508.
- Kendall, C., Isabelle, M., Bazant-hegemark, F., Hutchings, J., Orr, L., Babrah, J., and Stone, N. (2009). Vibrational spectroscopy : a clinical tool for cancer diagnostics. *Analyst*, 134: 1029–1045.
- Kenya, Globocan (2018). *International Agency for Research on Cancer*, 985: 1–2.
- Khanmohammadi, M., Rajabi, F. H., Garmarudi, A. B., and Mohammadzadeh, R. (2010). Chemometrics assisted investigation of variations in infrared spectra of blood samples obtained from women with breast cancer : a new approach for cancer diagnosis. *European Journal of Cancer Care*, 19: 352–359.
- Kivlighan, K. T., Granger, D. A., and Schwartz, E. B. (2005). Blood contamination and the measurement of salivary progesterone and estradiol. *Hormones and Behavior*, 47(3): 367–370.
- Klement, R. J., and Kämmerer, U. (2013). Is there a role for carbohydrate restriction in the treatment and prevention of cancer? *Clinical Nutrition: The Interface Between Metabolism, Diet, and Disease*, 8(1): 257–294.
- Kolehmainen, M., Martikainen, H., and Ruuskanen, J. (2001). Neural networks and periodic components used in air quality forecasting. *Atmospheric Environment*, 35: 815–825.

- Korir, A., Okerosi, N., Ronoh, V., Mutuma, G., and Parkin, M. (2015). Incidence of cancer in Nairobi , Kenya (2004 – 2008). *International Journal of Cancer*, 2059: 2053–2059.
- Krause, J. R. (2000). Morphology and classification of acute myeloid leukemias. *Clinics in Laboratory Medicine*, 20(1): 1–16.
- Larkin, P. (2011). *IR and Raman spectroscopy: Principles and spectral interpretation*. Elsevier Ltd. 89-97.
- Lasch, P. (2012). Spectral Pre-processing for Biomedical Vibrational Spectroscopy and Microspectroscopic Imaging. *Chemometrics and Intelligent Laboratory Systems*, 117: 100–114.
- Lasch, P., Waesche, W., Bindig, U., Naumann, D., and Mueller, G. J. (1997). Imaging of human colon carcinoma thin sections by FT-IR microspectrometry. *Optical Biopsies and Microscopic Techniques II*, 3197: 278–285.
- Lbany, A., and Ork, N. E. W. Y. (2015). *Multidimensional Raman Spectroscopic Signatures as a Tool for Forensic Identification of Body Fluid Traces : A Review*. 1223–1232.
- Leeman, M., Choi, J., Hansson, S., Storm, M. U., and Nilsson, L. (2018). Proteins and antibodies in serum , plasma , and whole blood — size characterization using asymmetrical flow field-flow fractionation (AF4). *Analytical and Bioanalytical Chemistry*, 410: 4867–4873.
- Leeuw, J. (2005). Modern Multidimensional Scaling: Theory and Applications. *Journal of Statistical Software*, 14(4): 2–4.
- Lewis, A. T., Gaifulina, R., Isabelle, M., Dorney, J., Woods, M. L., Lloyd, G. R., and Thomas, G. M. (2017). Mirrored stainless steel substrate provides improved signal for Raman spectroscopy of tissue and cells. *Journal of Raman Spectroscopy*, 48(1): 119–125.
- Lewis, I. R., and Edwards, H. G. M. (2001). *Handbook of Raman spectroscopy: From the Research Laboratory to the Process Line*. New York: Marcel Dekker Inc. 301-314.
- Li, X., Lin, J., Li, X., Yang, T., and Lin, J. (2012). Spectral analysis of human saliva for detection of lung cancer using surface- enhanced Raman spectroscopy. *Biomedical Optics*, 17(3): 1–5.
- Liu, P., Wen, Y., Huang, J., Xiong, A., Wen, J., Li, H., and Wu, R. (2019). A novel strategy of near-infrared spectroscopy dimensionality reduction for discrimination of grades, varieties and origins of green tea. *Vibrational Spectroscopy*, 105: 102984 -102989.
- Liu, W., Sun, Z., Chen, J., and Jing, C. (2016). Raman Spectroscopy in Colorectal Cancer Diagnostics : Comparison of PCA-LDA and PLS-DA Models. *Spectroscopy*, 64 (1): 1–7.
- Long, J., Zhang, C.-J., Zhu, N., Du, K., Yin, Y.-F., Tan, X., and Qin, L. (2018). Lipid metabolism

- and carcinogenesis, cancer development. *American Journal of Cancer Research*, 8(5): 778–791.
- Luo, Z., Wu, X., Guo, S., and Yee, B. (2008). Diagnosis of breast cancer tumor based on manifold learning and support vector machine. *International Conference on Information and Automation*, 703-707.
- Lyng, F. M., Traynor, D., Nguyen, T. N. Q., Meade, A. D., Rakib, F., Al-Saady, R., and Ali, M. H. (2019). Discrimination of breast cancer from benign tumours using Raman spectroscopy. *PLoS ONE*, 14(2): 1–13.
- Magalhães, F. L., Machado, A. M. C., Jr, E. P., Sahoo, S. K., Paula, A. M. De, Garcia, A. M., and Mamede, M. (2018). Raman spectroscopy with a 1064-nm wavelength laser as a potential molecular tool for prostate cancer diagnosis : a pilot study. *Biomedical Optics*, 23(12): 1–6.
- Malamud, D., and Rodriguez-Chavez, I. (2011). Saliva as a diagnostic fluid. *Dental Clinic in North America*, 55(1): 159–178.
- Marinello, P. C., Machado, K. L., Cecchini, R., and Cecchini, A. L. (2014). The Participation of Oxidative Stress in Breast Cancer Cells Progression and Treatment Resistance. *American Journal of Immunology*, 10(4): 207–2014.
- Marini, F., Bucci, R., Magri, A. L., and Magri, A. D. (2008). Artificial neural networks in chemometrics : History , examples and perspectives. *Microchemical Journal*, 88: 178–185.
- Martin, K. J., Fournier, M. V, Reddy, G. P. V., and Pardee, A. B. (2010). A need for basic Research on fluid-Based early detection biomarkers. *Cancer Research*, 70: 5203–5207.
- MartinEspinoza, J. C., GonzalezSolis, J. L., FraustoReyes, C., MirandaBeltran, M. L., and SoriaFregoso, C. (2008). Raman spectroscopy: A new proposal for the detection of leukemia using blood samples. *American Institute of Physics*, 1032: 252–254.
- Martinez, W. L., and Martinez, A. R. (2005). *Exploratory Data Analysis with MATLAB*®. London, United Kingdom: CRC Press. 42-46.
- Martins, L. D. O., and Junior, G. B. (2009). Detection of Masses in Digital Mammograms using K-means and Support Vector Machine. *Electronic Letters on Computer Vision and Image Analysis*, 8(2): 39–50.
- Masood, K., Rajpoot, N., Rajpoot, K., and Qureshi, H. (2006). Hyperspectral colon tissue classification using morphological analysis. *2nd International Conference on Emerging Technologies*, 13–14.
- Masters, J. R., and Stacey, G. N. (2007). Changing medium and passaging cell lines. *Nature*

Protocols, 2(9): 2276–2284.

- Matias, R., Silveira, L., Augusto, M., and Silva, R. S. (2011). Diagnostic model based on Raman spectra of normal, hyperplasia and prostate adenocarcinoma tissues in vitro. *Spectroscopy*, 25: 89–102.
- Matthäus, C., Bird, B., Miljkovi, M., Chernenko, T., Romeo, M., and Diem M. (2008). Infrared and Raman microscopy in cell biology. *Methods in Cell Biology*, 89: 275–308.
- McCreery, R. L. (2001). *Raman Spectroscopy for Chemical Analysis*. John Wiley & Sons, Ltd, 91-92.
- McLachlan, G. (1999). Mahalanobis Distance. *Resonance*, 2 (1): 20–26.
- Mcmenamy, R. H., Lund, C. C., and Lawrence, J. (1957). Unbound Amino Acid Concentrations in Human Blood Plasmas. *Clinical Investigation*, 36(12): 1672–1679.
- Meurman, J. H. (2010). Infectious and dietary risk factors of oral cancer. *Oral Oncology*, 46(6): 411–413.
- Mika, S., and Scho, B. (2002). Constructing Boosting Algorithms from SVMs : An Application to One-Class Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(9): 1184–1199.
- Ministry of Health, Kenya (2019). Kenya Cancer Policy 2019 - 2030. *Ministry of Health, Government of Kenya*, 7(2): 1–16.
- Mittal, S., Bansal, V., Garg, S., Atreja, G., and Bansal, S. (2011). The diagnostic role of saliva-A review. *Clinical and Experimental Dentistry*, 3(4): 314–320.
- Moisou, V., Stefanu U., Iancu, S. D., Moisoiu, T., Loga, L., Dican, L., Alesca, C. D., Boros, I., Jurj, A., Dima, D., Bgacean, C., Tetean, R., Burzo, E., Tomuleasa, C., Elec, F., and Leopold, N. (2019). SERS assessment of the cancer-specific methylation pattern of genomic DNA : towards the detection of acute myeloid leukemia in patients undergoing hematopoietic stem cell transplantation. *Analytical and Bioanalytical Chemistry*, 1: 1-7.
- Movasaghi, Z., Rehman, S., and Rehman, I. U. (2007). Raman Spectroscopy of Biological Tissues. *Applied Spectroscopy Reviews*, 42: 493–541.
- Mulligan, H. D., and Tisdale, M. J. (1991). Effect of the lipid-lowering agent bezafibrate on tumour growth rate in vivo. *British Journal of Cancer*, 64: 1035–1038.
- Musto, P., Calarco, A., Pannico, M., Manna, P. La, Margarucci, S., Tafuri, A., and Peluso, G. (2017). Hyperspectral Raman imaging of human prostatic cells : An attempt to differentiate normal and malignant cell lines by univariate and multivariate data analysis. *Spectrochimica*

- Acta Part A: Molecular and Biomolecular Spectroscopy*, 173: 476–488.
- Naghavi, M. (2015). The global burden of Cancer 2013. *JAMA Oncology*, 1(4): 505–527.
- Nargis, H. F., Nawaz, H., Ditta, A., Mahmood, T., Majeed, M., Rashid, N., and Bryne, H. (2019). Raman spectroscopy of blood plasma samples from breast cancer patients at different stages. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy* 222: 117210-117219.
- Ng, S.H., Shuib, A., Phang, S W., Sabri, M. I. A., and Muda, A. (2019). Development of blood mimicking fluid suspension using polymer particles. *Proceedings of the International Engineering Research Conference*, 2137 (2019): 1-8.
- Nguyen, C. T., Nguyen, J. T., Rutledge, S., Zhang, J., Wang, C., and Walker, G. C. (2010). Detection of chronic lymphocytic leukemia cell surface markers using surface enhanced Raman scattering gold nanoparticles. *Cancer Letters*, 292(1): 91-97.
- Nijssen, A., Schut, T. C. B., Heule, F., Caspers, P. J., Hayes, D. P., Neumann, M. H. A., and Puppels, G. J. (2002). Discriminating Basal Cell Carcinoma from its Surrounding Tissue by Raman Spectroscopy. *Investigative Dermatology*, 119: 64–69.
- Okonda, J. J., Angeyo, K. H., Mangala, J. M., and Kisia, S. M. (2017). A nested multivariate chemometrics based calibration strategy for direct trace biometal analysis in soft tissue utilizing Energy Dispersive X-Ray Fluorescence (EDXRF) and scattering spectrometry. *Applied Radiation and Isotopes*, 129: 49–56.
- Ong, Y. H., Lim, M., and Liu, Q. (2012). Comparison of principal component analysis and biochemical component analysis in Raman spectroscopy for the discrimination of apoptosis and necrosis in K562 leukemia cells. *Optics Express*, 20(20): 22158–22171.
- Panchbhai, A. S. (2012). Correlation of Salivary Glucose Level with Blood Glucose Level in Diabetes Mellitus. *Oral and Maxillofacial Research*, 3(3): 1–7.
- Parawira, S. (2009). Classification of hyperspectral breast images for cancer detection. *International conference on Chemometrics and Imaging*, 6 (7): 2–6.
- Patel, I. I., and Martin, F. L. (2010). Discrimination of zone-specific spectral signatures in normal human prostate using Raman spectroscopy. *Analyst*, 135: 3060–3069.
- Pelletier, M. j. (2003). Quantitative analysis using Raman spectrometry. *Applied Spectroscopy*, 57(1): 20A-42A.
- Pernot, E., Cardis, E., and Badie, C. (2014). Usefulness of saliva samples for biomarker studies in radiation research. *Cancer Epidemiology Biomarkers and Prevention*, 23(12): 2673–2680.
- Pfaffe, T., Cooper-white, J., Beyerlein, P., Kostner, K., and Punyadeera, C. (2011). Diagnostic

- Potential of Saliva : Current State and Future Applications. *Clinical Chemistry*, 57(5): 675–687.
- 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. (2007). Raman spectroscopy and multivariate analysis of serum samples from breast cancer patients. *Lasers in Medical Science*, 22: 229–236.
- Poehls, U. G., Hack, C. C., Ekici, A. B., Beckmann, M. W., Fasching, P. A., Ruebner, M., & Huebner, H. (2018). Saliva samples as a source of DNA for high throughput genotyping: An acceptable and sufficient means in improvement of risk estimation throughout mammographic diagnostics. *European Journal of Medical Research*, 23(1): 1–10.
- Porto-Mascarenhas, E. C., Assad, D. X., Chardin, H., Gozal, D., De Luca Canto, G., Acevedo, A. C., and Guerra, E. N. S. (2017). Salivary biomarkers in the diagnosis of breast cancer: A review. *Critical Reviews in Oncology and Hematology*, 110: 62–73.
- Qiu, S., Xu, Y., Huang, L., Zheng, W. E. I., Huang, C., and Huang, S. (2016). Non-invasive detection of nasopharyngeal carcinoma using saliva surface-enhanced Raman spectroscopy. *Oncology Letters*, 2: 884–890.
- Rao, P. V., Reddy, A. P., Lu, X., Dasari, S., Krishnaprasad, A., Biggs, E., and Nagalla, S. R. (2009). Proteomic identification of salivary biomarkers of type-2 diabetes. *Journal of Proteome Research*, 8(1): 239–245.
- Rasi, S., Brusca, A., Rinaldi, A., Cresta, S., Fangazio, M., de Paoli, L., and Rossi, D. (2011). Saliva is a reliable and practical source of germline DNA for genome-wide studies in chronic lymphocytic leukemia. *Leukemia Research*, 35(10): 1419–1422.
- Ratle, F., Camps-valls, G., and Weston, J. (2010). Semisupervised neural networks for efficient hyperspectral image classification. *IEEE Transactions On Geoscience And Remote Sensing*, 48(5): 2271–2282.
- Rauch, C., Feifel, E., Amann, E.-M., Spötl, H., Schennach, H., Pfaller, W., and Gstraunthaler, G. (2009). Alternatives to the use of Fetal Bovine Serum : Human platelet lysates as a serum substitute in cell culture media. *Altex* 28, 4(11): 305–316.
- Rehman, I., Movasaghi, Z., and Rehman, S. (2013). *Vibrational spectroscopy for tissue analysis*, CRC Press. 214-287.
- Rehman, S., Movasaghi, Z., Darr, J. A., and Rehman, I. U. (2010). Fourier transform infrared spectroscopic analysis of breast cancer tissues; Identifying differences between normal breast, invasive ductal carcinoma, and ductal carcinoma in situ of the breast. *Applied Spectroscopy*

- Reviews*, 45(5): 355–368.
- Rezaeianzadeh, M., Tabari, H., Arabi Yazdi, A., Isik, S., and Kalin, L. (2014). Flood flow forecasting using ANN, ANFIS and regression models. *Neural Computing and Applications*, 25(1): 25–37.
- Rodionova, O. Y., and Pomerantsev, A. L. (2006). Chemometrics : achievements and prospects. *Russian Chemical Review*, 75: 271–287.
- Ryabchykov, O., Guo, S., and Bocklitz, T. (2019). Analyzing Raman spectroscopic data. *Physical Sciences Reviews*, 4(2): 1–16.
- Saeyns, W., Mouazen, A. M., and Ramon, H. (2005). Potential for onsite and online analysis of pig manure using visible and near infrared reflectance spectroscopy. *Biosystems Engineering*, 91(4): 393–402.
- Sajda, P. (2006). Machine Learning for Detection and Diagnosis of Disease. *Annual Review in Biomedical Engineering*, 8: 537–565.
- Salem, O., Guerassimov, A., Mehaoua, A., Marcus, A., and Furht, B. (2014). Anomaly detection in medical wireless sensor networks using SVM and linear regression models. *International Journal of E-Health and Medical Communications*, 5(1): 20–45.
- Saroch, G., and Paul. M.P., R. (2012). A comparative study on UV spectrophotometric quantification of DNA extracted from human saliva. *Egyptian Journal of Forensic Sciences*, 2(4): 123–125.
- Schie, I. W. (2013). Methods and applications of Raman microspectroscopy to single -cell analysis. *Applied Spectroscopy*, 67(8): 813–828.
- Schrader, B. (1995). *Infrared and Raman spectroscopy: Methods and Applications*. Weinheini: 70-92..
- Schwabl, F. (2007). *Quantum Mechanics* , 4th Edition, New York: Berlin, Springer. 203-298.
- Scott, D. A., Renaud, D. E., Krishnasamy, S., Meriç, P., Buduneli, N., and Çetinkalp, S. (2010). Diabetes-related molecular signatures in infrared spectra of human saliva, *Bionalytics*, 2(3): 1–9.
- Semleit, D., Trampe, A., and Fissan, H. (1997). Fluctuations and noise of the optical output power of laser diodes and the effect on optical particle size determination. *Aerosol Science and Technology*, 26(4): 356–367.
- Shafer-peltier, K. E., Haka, A. S., Fitzmaurice, M., Crowe, J., Myles, J., Dasari, R. R., and Feld, M. S. (2002). Raman microspectroscopic model of human breast tissue : implications for breast

- cancer diagnosis in vivo. *Raman Spectroscopy*, 33: 552–563.
- Sheng, D., Liu, X., Li, W., Wang, Y., Chen, X., and Wang, X. (2013). Distinction of leukemia patients' and healthy persons' serum using FTIR spectroscopy. *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 101: 228–232.
- Shipp, D.W., Sinjab, F., and Notingher, I. (2017). Raman spectroscopy: techniques and applications in the life sciences. *Advances in Optics and Photonics*, 9(2): 315-390.
- Shirtcliff, E. A., Granger, D. A., Schwartz, E. B., Curran, M. J., Booth, A., and Overman, W. H. (2000). Assessing estradiol in biobehavioral studies using saliva and blood spots: Simple radioimmunoassay protocols, reliability, and comparative validity. *Hormones and Behavior*, 38(2): 137–147.
- Sichangi, E. K., Angeyo, H. K., Dehayem-Kamadjeu, A., and Mangala, M. (2018). Hybridized robust chemometrics approach for direct rapid determination of trace biometals in tissue utilizing energy dispersive X-ray fluorescence and scattering (EDXRFS) spectrometry. *Radiation Physics and Chemistry*, 153: 198–207.
- Sikirzhytski, V., Sikirzhytskaya, A., Lednev, I. K., Lbany, A., and Ork, N. E. (2011). Multidimensional Raman spectroscopic signatures as a tool for forensic identification of body fluid traces: A review. *Applied Spectroscopy*, 65(11): 1223–1232.
- Sikirzhytski, V., Virkler, K., and Lednev, I. K. (2010). Identification for Forensic Purposes. *Sensors*, 10: 2869–2884.
- Simeonova, P., Lovchinov, V., Dimitrov, D., and Radulov, I. (2010). Environmetric approaches for lake pollution assessment. *Environmental Monitoring and Assessment*, 164: 233-248.
- Singh, S., Gupta, V., Vij, R., Sharma, B., and Agarwal, R. (2014). Saliva : Diagnostic tool of Future. *Oral Pathology*, 6(5): 37–38.
- Singla, R., Chambayil, B., Khosla, A., and Santosh, J. (2011). Comparison of SVM and ANN for classification of eye events in EEG. *Biomedical Science and Engineering*, 4: 62–69.
- Smith, E., and Dent, G. (2005). *Modern Raman Spectroscopy – A Practical Approach*. John Wiley & Sons, Ltd, 503-505.
- Sobin, L. H., Gospodarowicz, and Wittekind, C. (2009). *TNM Classification of Malignant Tumours (Seventh Ed)*. A John Wiley and Sons Ltd, 1-302.
- Stone, N., Consuelo, M., Prieto, H., Crow, P., Uff, J., and Ritchie, A. W. (2007). The use of Raman spectroscopy to provide an estimation of the gross biochemistry associated with urological pathologies. *Analytical Bioanalytical Chemistry*, 387: 1657–1668.

- Streckfus, C, and Bigler, L. (2004). The use of soluble, salivary c-erbB-2 for the detection and post-operative follow-up of breast cancer in women: The results of a five-year translational research study. *Saliva-/Oral-Fluid-Based Diagnostic Markers of Disease, Biology*, 2(5): 17–24.
- Streckfus, Charles, Bigler, L., Dellinger, T., Dai, X., Kingman, A., and Thigpen, J. T. (2000). The Presence of Soluble c- erb B-2 in Saliva and Serum among Women with Breast Carcinoma : A Preliminary Study. *Clinical Cancer Research*, 6: 2363–2370.
- Suárez Sánchez, A., García Nieto, P. J., Riesgo Fernández, P., del Coz Díaz, J. J., and Iglesias-Rodríguez, F. J. (2011). Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). *Mathematical and Computer Modelling*, 54(5–6): 1453–1466.
- Suhandy, D., Suzuki, T., Ogawa, Y., Kondo, N., Naito, H., Ishihara, T. and Liu, W. (2012). A quantitative study for determination of glucose concentration using attenuated total reflectance terahertz (ATR-THz) spectroscopy. *Engineering in Agriculture, Environment and Food*, 5(3): 90–95.
- Sullivan, G. M., and Feinn, R. (2012). Using Effect Size-or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, 4(3): 279–282.
- Switonski, A., Michalak, M., Josinski, H., and Wojciechowski, K. (2010). Detection of tumor tissue based on the multispectral imaging. *ICCVG*, 2(6375): 325–333.
- Tabak, L. A. (2001). A revolution in biomedical assessment: the development of salivary diagnostics. *Journal of Dental Education*, 65(12): 1335–1339.
- Talari, A. C. S., Evans, C. A., Holen, I., and Coleman, R. E. (2015). Raman spectroscopic analysis differentiates between breast cancer cell lines. *Raman Spectroscopy*, 46(5): 421–427.
- Taleb, A., Diamond, J., Mcgarvey, J. J., Beattie, J. R., Toland, C., and Hamilton, P. W. (2006). Raman Microscopy for the Chemometric Analysis of Tumor Cells. *Physical Chemistry B*, 110: 19625–19631.
- Taleuzzaman, M. (2018). Limit of Blank (LOB), Limit of Detection (LOD), and Limit of Quantification (LOQ). *Organic and Medicinal Chemistry*, 7(5): 1–5.
- Tang, J., Rangayyan, R. M., Xu, J., Naqa, I. El, and Yang, Y. (2009). Computer-aided detection and diagnosis of breast cancer with mammography : Recent Advances. *IEEE Transactions On Information Technology In Biomedicine*, 13(2): 236–251.
- Thalheimer, W., and Cook, S. (2002). How to calculate effect sizes from published research : A simplified methodology. *Work-Learning Research*, 1: 1-9.

- Theophilou, G., Lima, K. M. G., Briggs, M., Martin-hirsch, P. L., Stringfellow, H. F., and Martin, F. L. (2015). A biospectroscopic analysis of human prostate tissue obtained from different time periods points to a trans-generational alteration in spectral phenotype. *Scientific Reports*, 5(13465): 1–14.
- Trauth, M. H. (2015). MATLAB[®] recipes for earth sciences: Fourth edition. Springer-Verlag: 375-411.
- Tu, J. V. (1996). Advantages and Disadvantages of Using Artificial Neural Networks versus Logistic Regression for Predicting Medical Outcomes. *Journal of Clinical Epidemiology*, 49(11): 1225–1231.
- Valderrama, P., Braga, J. W. B., and Poppi, R. J. (2007). Variable selection, outlier detection, and figures of merit estimation in a partial least-squares regression multivariate calibration model. A case study for the determination of quality parameters in the alcohol industry by near-infrared spectroscopy. *Journal of Agricultural and Food Chemistry*, 55(21): 8331–8338.
- Vanna, R., Tresoldi, C., Ronchi, P., Lenferink, A. T. M., Morasso, C., Mehn, D., and Gramatica, F. (2014). Raman spectroscopy for the assessment of acute myeloid leukemia: a proof of concept study. *Biomedical Vibrational Spectroscopy VI: Advances in Research and Industry*, 8939: 1-15.
- Vargas-Obieta, E., Martínez-Espinosa, J. C., Martínez-Zerega, B. E., Jave-Suárez, L. F., Aguilar-Lemarroy, A., and González-Solís, J. L. (2016). Breast cancer detection based on serum sample surface enhanced Raman spectroscopy. *Lasers in Medical Science*, 31(7): 1317–1324.
- Varmuza, K., and Filzmoser, P. (2008). *Introduction to Multivariate Statistical Analysis in Chemometrics*. Boca Raton: CRC Press. 40-100.
- Venkatachalam, P., Rao, L. L., Kumar, N. K., Jose, A., and Nazeer, S. S. (2008). Diagnosis of breast cancer based on FT-IR spectroscopy. *Perspectives in Vibrational Spectroscopy-American Institute of Physics*, 978: 144–148.
- Verma, P. (2017). Tip-Enhanced Raman Spectroscopy: Technique and Recent Advances. *Chemical Reviews*, 117: 6447-6466.
- Wachsmann-hogiu, S., Weeks, T., and Huser, T. (2009). Chemical analysis in vivo and in vitro by Raman spectroscopy — from single cells to humans. *Current Opinion in Biotechnology*, 20: 63-73.
- Walt, D. R., Blicharz, T. M., Hayman, R. B., Rissin, D. M., Bowden, M., Siqueira, W. L., and Brody, J. S. (2007). Microsensor arrays for saliva diagnostics. *Annals of the New York Academy*

- of Sciences*, 1098: 389–400.
- Wang, G., Ding, Q., and Hou, Z. (2008). Independent component analysis and its applications in signal processing for analytical chemistry. *Trends in Analytical Chemistry*, 27(4): 368–376.
- Wang, L., and Mizaikoff, B. (2008). Application of multivariate data-analysis techniques to biomedical diagnostics based on mid-infrared spectroscopy. *Analytical Bioanalytical Chemistry*, 391: 1641–1654.
- Wang, W., Zhao, J., Short, M., and Zeng, H. (2014). Real-time in vivo cancer diagnosis using Raman spectroscopy. *Biophotonics*, 19: 1–19.
- Weinberg, S. (1991). An Introduction to Multidimensional IRT. *Measurement and Evaluation in Counseling and Development*, 24: 12–36.
- Weinmann, P., Jouan, M., Dao, N. Q., Lacroix, B., Groiselle, C., Bonte, J., Luc, G. (1998). Quantitative analysis of cholesterol and cholesteryl esters in human atherosclerotic plaques using near-infrared Raman spectroscopy. *Atherosclerosis*, 140(1): 81–88.
- Wu, W., Gong, H., Liu, M., Chen, G., and Chen, R. (2015). Noninvasive Breast Tumors Detection based on Saliva Protein Surface Enhanced Raman Spectroscopy and Regularized Multinomial Regression. *8th International Conference on BioMedical Engineering and Informatics*, 214–218.
- Wythoff, B. J. (1993). Backpropagation neural networks. A tutorial. *Chemometrics and Intelligent Laboratory Systems*, 18(2): 115–155.
- Xuchun, L., Wang, L., and Eric, S. (2007). AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence*, 21: 785–795.
- Yang, T., Guo, S., Wu, X., and Wu, X. (2007). An Approach Based on Immune Algorithm and SVM for Detection and Classification of Microcalcifications. *IEEE Proceedings*, 30670538: 1–4.
- Yao, F., Coquery, J., and Lê Cao, K. A. (2012). Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. *BMC Bioinformatics*, 13(24): 1–15.
- Yu, Y., Lin, J., Lin, D., Feng, S., Chien, W., Huang, Z., Huang H., and Chen, A. (2017). Leukemia cells detection based on electroporation assisted surface-enhanced Raman scattering. *Biomedical Optical Express*, 8(9): 4108-4121.
- Zhang, C., Cheng, X., Li, J., Zhang, P., Yi, P., Xu, X., and Zhou, X. (2016). Saliva in the diagnosis

- of diseases. *International Journal of Oral Science*, 8(3): 133–137.
- Zhang, Y., Sun, J., Lin, C., Abemayor, E., Wang, M. B., and Tw, D. (2014). The Emerging Landscape of Salivary Diagnostics. *OHDM*, 13(2): 200–210.
- Zhao, J., Short, M., Braun, T., Lui, H., McLean, D., and Zeng, H. (2014). Clinical Raman measurements under special ambient lighting illumination. *Journal of Biomedical Optics*, 19(11): 1-2.
- Zhou, P. (2015). Choosing the Most Suitable Laser Wavelength For Your Raman Application. *Raman Spectroscopy*, 1(855): 1–6.

Appendix I

Clinical diagnosis of malignant breast tumor patients and healthy subjects

| | CHT | Malignant breast tumor (<i>n</i> = 20) | Healthy / control subjects (<i>n</i> = 23) |
|-----------------|-----|---|---|
| Mean age, years | | 59 | 51 |
| Carcinoma | | | |
| Stage 1 | No | 3 | NA |
| Stage 2 | No | 7 | NA |
| Stage 3 | No | 10 | NA |

Abbreviations: *n*, number; NA, not applicable; CHT, chemotherapy cancer treatment
(No, none on treatment)

Appendix 11

Clinical diagnosis of leukemia patients and healthy subjects

| | CHT | Leukemia (<i>n</i> = 9) | Healthy subjects / control (<i>n</i> = 18) |
|-----------------|-----|-----------------------------|--|
| Mean age, years | | 66 | 42 |
| Carcinoma | | | |
| Stage 1 | No | NA | NA |
| Stage 2 | No | NA | NA |
| Stage 3 | No | 9 | NA |

Abbreviations: *n*, number; NA, not applicable; CHT, chemotherapy cancer treatment
(No, none on treatment)

Appendix III

Principal Component Analysis – Linear Discriminant Analysis (PCA-LDA)

```
// Variables
// X                dataset [samples x variables]
// class            class vector
// cv_type           type of cross validation
// cv_groups         number of cross validation groups
// class_prob        prior probability
// method           'linear' (LDA) or 'quadratic' (QDA)
// pret_type         data pretreatment
// max_comp          maximum number of components to be calculated

// data pretreatment
a = mean(X);
s = std(X);
m = min(X);
M = max(X);
if strcmp(pret_type,'cent')
    amat = repmat(a,size(X,1),1);
    X_scal = X - amat;
elseif strcmp(pret_type,'scal')
    f = find(s>0);
    smat = repmat(s,size(X,1),1);
    X_scal = zeros(size(X,1),size(X,2));
    X_scal = X(:,f)./smat(:,f);
elseif strcmp(pret_type,'auto')
    f = find(s>0);
    amat = repmat(a,size(X,1),1);
    smat = repmat(s,size(X,1),1);
    X_scal = zeros(size(X,1),size(X,2));
    X_scal(:,f) = (X(:,f) - amat(:,f))./smat(:,f);
```



```

elseif strcmp(pret_type,'rang')
    f = find(M - m > 0);
    mmat = repmat(m,size(X,1),1);
    Mmat = repmat(M,size(X,1),1);
    X_scal = zeros(size(X,1),size(X,2));
    X_scal(:,f) = (X(:,f) - mmat(:,f))./(Mmat(:,f) - mmat(:,f));
else
    X_scal = X;
end
param.a = a;
param.s = s;
param.m = m;
param.M = M;
param.pret_type = pret_type;
// selection of the optimal number of components for PCA coupled with DA by means of
// cross-validation
function res = dacompSel(X,class,cv_type,cv_groups,class_prob,method,pret_type,max_comp)
[n,p] = size(X);
r = min(n,p);
if r > max_comp
    r = max_comp;
end
hwait = waitbar(0,'cross validating models','CreateCancelBtn','setappdata(gcf,'canceling',1));
setappdata(hwait,'canceling',0);
for k = 1:r
    if ~ishandle(hwait)
        res.er = NaN;
        res.ner = NaN;
        res.not_ass = NaN;
        break
    elseif getappdata(hwait,'canceling')
        res.er = NaN;

```

```

    res.ner = NaN;
    res.not_ass = NaN;
    break
else
    waitbar(k/r)
    out = dacv(X,class,cv_type,cv_groups,class_prob,method,k,pret_type);
    res.er(k) = out.class_param.er;
    res.ner(k) = out.class_param.ner;
    res.not_ass(k) = out.class_param.not_ass;
end
end
if ishandle(hwait)
    delete(hwait)
end
res.settings.pret_type = pret_type;
res.settings.cv_type = cv_type;
res.settings.cv_groups = cv_groups;
res.settings.class_prob = class_prob;
res.settings.method = method;
// cross validation for Discriminant Analysis (DA)
function cv = dacv(X,class,cv_type,cv_groups,class_prob,method,num_comp,pret_type)
if iscell(class)
    class_string = class;
    [class,class_labels] = calc_class_string(class_string);
else
    class_string = {};
    class_labels = {};
end
nobj = size(X,1);
if strcmp(cv_type,'boot')
    hwait = waitbar(0,'bootstrap validation');
    out_bootstrap = zeros(nobj,1);

```

```

class_pred = [];
class_true = [];
for i=1:cv_groups
    waitbar(i/cv_groups)
    out = ones(nobj,1);
    whos_in = [];
    for k=1:nobj
        r = ceil(rand*nobj);
        whos_in(k) = r;
    end
    out(whos_in) = 0;
    // counters for left out samples
    boot_how_many_out(i)=length(find(out == 1));
    out_bootstrap(find(out == 1)) = out_bootstrap(find(out == 1)) + 1;
    Xext = X(find(out == 1),:);
    class_ext = class(find(out == 1));
    Xtrain = X(whos_in,:);
    class_train = class(whos_in);
    if nargin == 8
        model = dafit(Xtrain,class_train,class_prob,method,num_comp,pret_type);
    else
        model = dafit(Xtrain,class_train,class_prob,method);
        num_comp = 0;
        pret_type = 'none';
    end
    pred = dapred(Xext,model);
    class_pred = [class_pred; pred.class_pred];
    class_true = [class_true; class_ext];
end
class = class_true;
delete(hwait);
elseif strcmp(cv_type,'rand')

```

```

hwait = waitbar(0,'montecarlo validation');
out_rand = zeros(nobj,1);
perc_in = 0.8;
take_in = round(nobj*perc_in);
class_pred = [];
class_true = [];
for i=1:cv_groups
    waitbar(i/cv_groups)
    out = ones(nobj,1);
    whos_in = randperm(nobj);
    whos_in = whos_in(1:take_in);
    out(whos_in) = 0;
    % counters for left out samples
    out_rand(find(out == 1)) = out_rand(find(out == 1)) + 1;
    Xext = X(find(out == 1),:);
    class_ext = class(find(out == 1));
    Xtrain = X(whos_in,:);
    class_train = class(whos_in);
    if nargin == 8
        model = dafit(Xtrain,class_train,class_prob,method,num_comp,pret_type);
    else
        model = dafit(Xtrain,class_train,class_prob,method);
        num_comp = 0;
        pret_type = 'none';
    end
    pred = dapred(Xext,model);
    class_pred = [class_pred; pred.class_pred];
    class_true = [class_true; class_ext];
end
class = class_true;
delete(hwait);
else

```

```

class_pred = zeros(size(X,1),1);
obj_in_block = fix(nobj/cv_groups);
left_over = mod(nobj,cv_groups);
st = 1;
en = obj_in_block;
for i = 1:cv_groups
    in = ones(size(X,1),1);
    if strcmp(cv_type,'vene') % venetian blinds
        out = [i:cv_groups:nobj];
    else % contiguous blocks
        if left_over == 0
            out = [st:en];
            st = st + obj_in_block; en = en + obj_in_block;
        else
            if i < cv_groups - left_over
                out = [st:en];
                st = st + obj_in_block; en = en + obj_in_block;
            elseif i < cv_groups
                out = [st:en + 1];
                st = st + obj_in_block + 1; en = en + obj_in_block + 1;
            else
                out = [st:nobj];
            end
        end
    end
end
in(out) = 0;
Xtrain = X(find(in==1),:);
class_train = class(find(in==1));
Xext = X(find(in==0),:);
if nargin == 8
    model = dafit(Xtrain,class_train,class_prob,method,num_comp,pret_type);
else

```

```

        model = dafit(Xtrain,class_train,class_prob,method);
        num_comp = 0;
        pret_type = 'none';
    end
    pred = dapred(Xext,model);
    class_pred(find(in==0)) = pred.class_pred;
end
end
class_param = calc_class_param(class_pred,class);
cv.class_pred = class_pred;
if length(class_labels) > 0
    cv.class_pred_string = calc_class_string(cv.class_pred,class_labels);
end
cv.class_param = class_param;
cv.settings.cv_groups = cv_groups;
cv.settings.cv_type = cv_type;
cv.settings.class_prob = class_prob;
cv.settings.method = method;
cv.settings.num_comp = num_comp;
cv.settings.pret_type = pret_type;
// fit Discriminant Analysis (DA)
function model = dafit(X,class,class_prob,method,num_comp,pret_type)
if iscell(class)
    class_string = class;
    [class,class_labels] = calc_class_string(class_string);
else
    class_string = {};
    class_labels = {};
end
n = size(X,1);
nclass = max(class);
if class_prob == 2

```

```

for g = 1:nclass
    obj_cla(g) = length(find(class == g));
end
prob = obj_cla/n;
else
    prob = NaN;
end
if nargin == 6
    modelpca = pca_model(X,num_comp,pret_type);
    Xtrain = modelpca.T;
else
    Xtrain = X;
    modelpca = NaN;
    num_comp = 0;
    pret_type = 'none';
end
// if linear and not with PCs check for pooled estimate of covariance
doit = 1;
if strcmp('linear',method) & nargin < 6
    doit = pec(X,class);
end
if doit
    % fitting
    if class_prob == 1
        [class_calc,e,prob_calc] = classify(Xtrain,Xtrain,class,method);
    else
        [class_calc,e,prob_calc] = classify(Xtrain,Xtrain,class,method,prob);
    end
    % calculates canonical variables for lda
    if strcmp('linear',method)
        class_unfold = zeros(size(Xtrain,1),max(class)-1);
        for g=1:max(class)-1

```

```

        class_unfold(find(class==g),g) = 1;
    end
    [L,B,r,S,V] = canoncorr(Xtrain,class_unfold);
    for k=1:size(L,1);
        for j=1:size(L,2)
            Lstd(k,j) = L(k,j)*std(Xtrain(:,k));
        end
    end
end
end
else
    class_calc = ones(size(X,1),1);
    L = zeros(size(X,2),1);
    Lstd = zeros(size(X,2),1);
    S = zeros(size(X,1),1);
end
class_param = calc_class_param(class_calc,class);
if strcmp(method,'linear')
    if num_comp > 0
        model.type = 'pcalda';
    else
        model.type = 'lda';
    end
else
    if num_comp > 0
        model.type = 'pcaqda';
    else
        model.type = 'qda';
    end
end
model.class_calc = class_calc;
if length(class_labels) > 0
    model.class_calc_string = calc_class_string(model.class_calc,class_labels);
end

```



```

end
model.prob = prob_calc;
model.class_param = class_param;
if strcmp('linear',method)
    model.L = L;
    model.Lstd = Lstd;
    model.S = S;
end
model.settings.pret_type = pret_type;
model.settings.class_prob = class_prob;
model.settings.prob = prob;
model.settings.method = method;
model.settings.modelpca = modelpca;
model.settings.num_comp = num_comp;
model.settings.raw_data = X;
model.settings.class = class;
model.cv = [];
model.labels.variable_labels = {};
model.labels.sample_labels = {};
model.labels.class_labels = class_labels;

function doit = pec(X,class)
for g = 1:max(class)
    gmeans(g,:) = mean(X(find(class == g),:));
end
// Pooled estimate of covariance
[Q,R] = qr(X - gmeans(class,:), 0);
R = R / sqrt(size(X,1) - max(class)); % SigmaHat = R'*R
s = svd(R);
if any(s <= eps^(3/4)*max(s))
    doit = 0;

```

```

disp('The pooled covariance matrix of training samples must be positive definite. model not
calculated');
else
    doit = 1;
end
// prediction with Discriminant Analysis (DA)
function pred = dapred(X,model)
class_prob = model.settings.class_prob;
method = model.settings.method;
Xtrain = model.settings.raw_data;
num_comp = model.settings.num_comp;
class_train = model.settings.class;
prob = model.settings.prob;
if num_comp > 0
    modelpca = pca_project(X,model.settings.modelpca);
    Xin = modelpca.Tpred;
    Xtrain = modelpca.T;
else
    Xin = X;
end
// prediction
if class_prob == 1
    [class_pred,e,prob_pred] = classify(Xin,Xtrain,class_train,method);
else
    [class_pred,e,prob_pred] = classify(Xin,Xtrain,class_train,method,prob);
end
// prediction of scores on canonical variables only for LDA
if strcmp('linear',method)
    [a,param] = data_pretreatment(Xtrain,'cent');
    Xin_cent = test_pretreatment(Xin,param);
    pred.S = Xin_cent*model.L;
end

```

```
pred.class_pred = class_pred;
if length(model.labels.class_labels) > 0
    pred.class_pred_string = calc_class_string(pred.class_pred,model.labels.class_labels);
end
pred.prob = prob_pred;
if num_comp > 0
    pred.T = modelpca.Tpred;
    pred.modelpca = modelpca;
end
```

Appendix IV

Independent component analysis (FASTICA_version)

// Basic parameters in fixed-point algorithm:

```
// 'approach'          decorrelation
// 'numOfIC'          number of independent components
// Linearity          value of 'g':  Nonlinearity used:
                        'pow3' (default)  g(u)=u^3
                        'tanh'          g(u)=tanh(a1*u)
                        'gauss'         g(u)=u*exp(-a2*u^2/2)
                        'skew'         g(u)=u^2

// 'finetune'        on, off
// 'mu'              Step size = 1.
// 'stabilization'   Values 'on' or 'off'
```

// --Controlling convergence

```
// 'epsilon'          (number) stopping criterion. Default is 0.0001
// 'maxNumIterations' maximum number of iterations = 1000
// 'maxFinetune'      maximum number of iterations = 100
// 'sampleSize'       (number) [0 - 1] %
// 'initGuess'        default is random
```

// Check some basic requirements of the data

```
if nargin == 0,
    error ('You must supply the mixed data as input argument. ');
end

if length (size (mixedsig)) > 2,
    error ('Input data can not have more than two dimensions. ');
end

if any (any (isnan (mixedsig))),
    error ('Input data contains NaN"s. ');
end

if ~isa (mixedsig, 'double')
    fprintf ('Warning: converting input data into regular (double) precision.\n');
```

```

    mixedsig = double (mixedsig);
end
// Remove the mean and check the data
[mixedsig, mixedmean] = remmean(mixedsig);
[Dim, NumOfSamp] = size(mixedsig);
% Default values for optional parameters
verbose      = 'on';
// Default values for 'pcamat' parameters
firstEig     = 1;
lastEig      = Dim;
interactivePCA = 'off';
// Default values for 'fpica' parameters
approach     = 'defl';
numOfIC      = Dim;
g            = 'pow3';
finetune     = 'off';
a1           = 1;
a2           = 1;
myy         = 1;
stabilization = 'off';
epsilon      = 0.0001;
maxNumIterations = 1000;
maxFinetune  = 5;
initState    = 'rand';
guess        = 0;
sampleSize   = 1;
displayMode  = 'off';
displayInterval = 1;

// Parameters for fastICA - i.e. this file
b_verbose = 1;
jumpPCA = 0;

```

```

jumpWhitening = 0;
only = 3;
userNumOfIC = 0;
// Read the optional parameters
if (rem(length(varargin),2)==1)
    error('Optional parameters should always go by pairs');
else
    for i=1:2:(length(varargin)-1)
        if ~ischar (varargin{i}),
            error (['Unknown type of optional parameter name (parameter' ...
                ' names must be strings).']);
        end
        % change the value of parameter
        switch lower (varargin{i})
            case 'stabilization'
                stabilization = lower (varargin{i+1});
            case 'maxfinetune'
                maxFinetune = varargin{i+1};
            case 'samplesize'
                sampleSize = varargin{i+1};
            case 'verbose'
                verbose = lower (varargin{i+1});
                if strcmp (verbose, 'off'), b_verbose = 0; end
            case 'firsteig'
                firstEig = varargin{i+1};
            case 'lasteig'
                lastEig = varargin{i+1};
            case 'interactivepca'
                interactivePCA = lower (varargin{i+1});
            case 'approach'
                approach = lower (varargin{i+1});
            case 'numofic'

```

```

numOfIC = varargin{i+1};
// User has supplied new value for numOfIC.
userNumOfIC = 1;
case 'g'
g = lower(varargin{i+1});
case 'finetune'
finetune = lower(varargin{i+1});
case 'a1'
a1 = varargin{i+1};
case 'a2'
a2 = varargin{i+1};
case {'mu', 'myy'}
myy = varargin{i+1};
case 'epsilon'
epsilon = varargin{i+1};
case 'maxnumiterations'
maxNumIterations = varargin{i+1};
case 'initguess'
// no use setting 'guess' if the 'initState' is not set
initState = 'guess';
guess = varargin{i+1};
case 'displaymode'
displayMode = lower(varargin{i+1});
case 'displayinterval'
displayInterval = varargin{i+1};
case 'pcae'

// calculate if there are enough parameters to skip PCA
jumpPCA = jumpPCA + 1;
E = varargin{i+1};
case 'pcad'
// calculate if there are enough parameters to skip PCA

```

```

jumpPCA = jumpPCA + 1;
D = varargin{i+1};
case 'whitesig'
    // calculate if there are enough parameters to skip PCA and whitening
    jumpWhitening = jumpWhitening + 1;
    whitesig = varargin{i+1};
case 'whitemat'
    % calculate if there are enough parameters to skip PCA and whitening
    jumpWhitening = jumpWhitening + 1;
    whiteningMatrix = varargin{i+1};
case 'dewhitemat'
    // calculate if there are enough parameters to skip PCA and whitening
    jumpWhitening = jumpWhitening + 1;
    dewhitematMatrix = varargin{i+1};
case 'only'
    // if the user only wants to calculate PCA or...
    switch lower(varargin{i+1})
        case 'pca'
            only = 1;
        case 'white'
            only = 2;
        case 'all'
            only = 3;
        end
    otherwise
        error(['Unrecognized parameter: "' varargin{i} '"']);
    end;
end;
end
// print information about data
if b_verbose
    fprintf('Number of signals: %d\n', Dim);

```



```

    fprintf('Number of samples: %d\n', NumOfSampl);
end
if Dim > NumOfSampl
    if b_verbose
        fprintf('Warning: ');
        fprintf('The signal matrix may be oriented in the wrong way.\n');
        fprintf('In that case transpose the matrix.\n\n');
    end
end
end
// Calculating PCA
if jumpWhitening == 3
    if b_verbose,
        fprintf('Whitened signal and corresponding matrices supplied.\n');
        fprintf('PCA calculations not needed.\n');
    end;
else
    if jumpPCA == 2,
        if b_verbose,
            fprintf('Values for PCA calculations supplied.\n');
            fprintf('PCA calculations not needed.\n');
        end;
    else
        if (jumpPCA > 0) & (b_verbose),
            fprintf('You must supply all of these in order to jump PCA:\n');
            fprintf('"pcaE", "pcaD".\n');
        end;
        [E, D]=pcamat(mixedsig, firstEig, lastEig, interactivePCA, verbose);
    end
end
end
// Whitening the data
    if jumpWhitening == 3,
        if b_verbose,

```

```

    fprintf('Whitening not needed.\n');
end;
else
if (jumpWhitening > 0) & (b_verbose),
    fprintf('You must supply all of these in order to jump whitening:\n');
    fprintf('"whiteSig", "whiteMat", "dewhiteMat".\n');
end;
[whitesig, whiteningMatrix, dewhiteningMatrix] = whitenv ...
    (mixedsig, E, D, verbose);
end
end % if only > 1
Dim = size(whitesig, 1);
if numOfIC > Dim
    numOfIC = Dim;
if (b_verbose & userNumOfIC)
    fprintf('Warning: estimating only %d independent components\n', numOfIC);
    fprintf('(Can't estimate more independent components than dimension of data)\n');
end
end
// Calculate the ICA with fixed point algorithm.
[A, W] = fpica (whitesig, whiteningMatrix, dewhiteningMatrix, approach, ...
    numOfIC, g, finetune, a1, a2, myy, stabilization, epsilon, ...
    maxNumIterations, maxFinetune, initState, guess, sampleSize, ...
    displayMode, displayInterval, verbose);
// Check for valid return
if ~isempty(W)
    % Add the mean back in.
if b_verbose
    fprintf('Adding the mean back to the data.\n');
end
icasig = W * mixedsig + (W * mixedmean) * ones(1, NumOfSampl);
%icasig = W * mixedsig;

```

```

if b_verbose & ...
    (max(abs(W * mixedmean)) > 1e-9) & ...
    (strcmp(displayMode,'signals') | strcmp(displayMode,'on'))
    fprintf('Note that the plots don"t have the mean added.\n');
end
else
    icasig = [];
end
end % if only > 2
if only == 1 % only PCA
    Out1 = E;
    Out2 = D;
elseif only == 2 % only PCA & whitening
    if nargout == 2
        Out1 = whiteningMatrix;
        Out2 = dwhiteningMatrix;
    else
        Out1 = whitesig;
        Out2 = whiteningMatrix;
        Out3 = dwhiteningMatrix;
    end
end
else % ICA
    if nargout == 2
        Out1 = A;
        Out2 = W;
    else
        Out1 = icasig;
        Out2 = A;
        Out3 = W;
    end
end
end

```

Appendix V

Partial least squares discriminant analysis (PLS-DA)

// Variables

```
// X          dataset [samples x variables]
// class      class vector, class labels can be
// pret_type  data pretreatment
// cv_type    type of cross validation
// cv_groups  number of cv groups
// assign_method  assignation method (bayes, max)
// model      plsda model calculated by means of plsdafit
// selection of the optimal number of latent variables for PLSDA by means of cross-validation
function res = plsdacompSel(X,class,pret_type,cv_type,cv_groups,assign_method)
[n,p] = size(X);
r = min(n,p);
if r > 20
    r = 20;
end
if r > 2
    r = r - 1;
end
hwait = waitbar(0,'cross validating models','CreateCancelBtn','setappdata(gcf,'canceling',1));
setappdata(hwait,'canceling',0);
for k = 1:r
    if ~ishandle(hwait)
        res.er = NaN;
        res.ner = NaN;
        res.not_ass = NaN;
        break
    elseif getappdata(hwait,'canceling')
        res.er = NaN;
        res.ner = NaN;
```

```

    res.not_ass = NaN;
    break
else
    waitbar(k/r)
    out = plsdcv(X,class,k,pret_type,cv_type,cv_groups,assign_method);
    res.er(k) = out.class_param.er;
    res.ner(k) = out.class_param.ner;
    res.not_ass(k) = out.class_param.not_ass;
end
end
if ishandle(hwait)
    delete(hwait)
end
res.settings.pret_type = pret_type;
res.settings.cv_type = cv_type;
res.settings.cv_groups = cv_groups;
res.settings.assign_method = assign_method;
// cross-validate Partial Least Squares Discriminant Analysis (PLSDA)
function cv = plsdcv(X,class,comp,pret_type,cv_type,cv_groups,assign_method)
if iscell(class)
    class_string = class;
    [class,class_labels] = calc_class_string(class_string);
else
    class_string = {};
    class_labels = {};
end
y = class;
x = X;
nobj=size(x,1);
if strcmp(cv_type,'boot')
    hwait = waitbar(0,'bootstrap validation');
    out_bootstrap = zeros(nobj,1);

```

```

assigned_class = [];
class_true = [];
for i=1:cv_groups
    waitbar(i/cv_groups)
    out = ones(nobj,1);
    whos_in = [];
    for k=1:nobj
        r = ceil(rand*nobj);
        whos_in(k) = r;
    end
    out(whos_in) = 0;
    % counters for left out samples
    boot_how_many_out(i)=length(find(out == 1));
    out_bootstrap(find(out == 1)) = out_bootstrap(find(out == 1)) + 1;
    x_out = x(find(out == 1),:);
    y_out = y(find(out == 1));
    x_in = x(whos_in,:);
    y_in = y(whos_in,:);
    model = plsdafit(x_in,y_in,comp,pret_type,assign_method,0);
    pred = plsdaPred(x_out,model);
    assigned_class = [assigned_class; pred.class_pred];
    class_true = [class_true; class(find(out == 1))];
    for g=1:size(y,2)
        rmsec(g) = NaN;
        quantitative_class(g) = NaN;
    end
end
class = class_true;
assigned_class = assigned_class';
delete(hwait);
elseif strcmp(cv_type,'rand')
    hwait = waitbar(0,'montecarlo validation');

```

```

assigned_class = [];
out_rand = zeros(nobj,1);
perc_in = 0.8;
take_in = round(nobj*perc_in);
class_true = [];
for i=1:cv_groups
    waitbar(i/cv_groups)
    out = ones(nobj,1);
    whos_in = randperm(nobj);
    whos_in = whos_in(1:take_in);
    out(whos_in) = 0;
    % counters for left out samples
    out_rand(find(out == 1)) = out_rand(find(out == 1)) + 1;
    x_out = x(find(out == 1),:);
    y_out = y(find(out == 1));
    x_in = x(whos_in,:);
    y_in = y(whos_in,:);
    model = plsdafit(x_in,y_in,comp,pret_type,assign_method,0);
    pred = plsdaPred(x_out,model);
    assigned_class = [assigned_class; pred.class_pred];
    class_true = [class_true; class(find(out == 1))];
    for g=1:size(y,2)
        rmsec(g) = NaN;
        quantitative_class(g) = NaN;
    end
end
class = class_true;
assigned_class = assigned_class';
delete(hwait);
else
    quantitative_class = zeros(nobj,max(class));
    class_pred = zeros(nobj,1);

```

```

obj_in_block = fix(nobj/cv_groups);
left_over = mod(nobj,cv_groups);
st = 1;
en = obj_in_block;
for i = 1:cv_groups
    in = ones(size(x,1),1);
    if strcmp(cv_type,'vene') % venetian blinds
        out = [i:cv_groups:nobj];
    else % contiguous blocks
        if left_over == 0
            out = [st:en];
            st = st + obj_in_block; en = en + obj_in_block;
        else
            if i < cv_groups - left_over
                out = [st:en];
                st = st + obj_in_block; en = en + obj_in_block;
            elseif i < cv_groups
                out = [st:en + 1];
                st = st + obj_in_block + 1; en = en + obj_in_block + 1;
            else
                out = [st:nobj];
            end
        end
    end
end
in(out) = 0;
x_in = x(find(in),:);
y_in = y(find(in),:);
x_out = x(find(in == 0),:);
model = plsdafit(x_in,y_in,comp,pret_type,assign_method,0);
out = plsdapred(x_out,model);
assigned_class(find(in == 0)) = out.class_pred;
quantitative_class(find(in == 0),:) = out.yc;

```



```

end
for g=1:size(y,2)
    C = calc_reg_param(y(:,g),quantitative_class(:,g));
    rmsec(g) = C.RMSEC;
end
end
class_param = calc_class_param(assigned_class',class);
cv.class_pred = assigned_class';
if length(class_labels) > 0
    cv.class_pred_string = calc_class_string(cv.class_pred,class_labels);
end
cv.class_param = class_param;
cv.yc = quantitative_class;
cv.rmsec = rmsec;
cv.settings.cv_groups = cv_groups;
cv.settings.cv_type = cv_type;
cv.settings.num_comp = comp;
cv.settings.pret_type = pret_type;
// assign samples for PLSDA on the basis of thresholds and calculated responses
function assigned_class = plsdafindclass(yc,class_thr)
nobj = size(yc,1);
nclass = size(yc,2);
for i = 1:nobj
    pred = yc(i,:);
    chk_ass = zeros(1,nclass);
    for c = 1:nclass
        if pred(c) > class_thr(c); chk_ass(c) = 1; end;
    end
    if length(find(chk_ass)) == 1
        assigned_class(i) = find(chk_ass);
    else
        assigned_class(i) = 0;
    end
end

```

```

    end
end
// find the class thresholds for PLSDA
function res = plsdafindthr(yc,class)
rsize = 100;
for g=1:size(yc,2)
    class_in = ones(size(class,1),1);
    class_in(find(class ~= g)) = 2;
    count = 0;
    y_in = yc(:,g);
    miny = min(y_in);
    thr = max(y_in);
    step = (thr - miny)/rsize;
    spsn = [];
    while thr > miny
        count = count + 1;
        class_calc_in = ones(size(class,1),1);
        thr = thr - step;
        sample_out_g = find(y_in < thr);
        class_calc_in(sample_out_g) = 2;
        cp = calc_class_param(class_calc_in,class_in);
        sp(count,g) = cp.specificity(1);
        sn(count,g) = cp.sensitivity(1);
        thr_val(count,g) = thr;
    end
end
end
% find best thr based on bayesian discrimination threshold
for g=1:max(class)
    P_g = yc(find(class==g),g);
    P_notg = yc(find(class~=g),g);
    m_g = mean(P_g); s_g = std(P_g);
    m_notg = mean(P_notg); s_notg = std(P_notg);

```

```

stp = abs(m_g - m_notg)/1000;
where = [m_notg:stp:m_g];
% fit normal distribution
% npdf_g = normpdf(where,m_g,s_g);
x_g = (where - m_g) ./ s_g;
npdf_g = exp(-0.5 * x_g.^2) ./ (sqrt(2*pi) .* s_g);
% npdf_notg = normpdf(where,m_notg,s_notg);
x_notg = (where - m_notg) ./ s_notg;
npdf_notg = exp(-0.5 * x_notg.^2) ./ (sqrt(2*pi) .* s_notg);
minval = NaN;
for k=1:length(where)
    diff = abs(npdf_g(k)-npdf_notg(k));
    if isnan(minval) || diff < minval
        minval = diff;
        class_thr(g) = where(k);
    end
end
if isnan(minval)
    class_thr(g) = mean([m_g m_notg]);
end
end
res.class_thr = class_thr;
res.sp = sp;
res.sn = sn;
res.thr_val = thr_val;
// fit Partial Least Squares Discriminant Analysis (PLSDA)
function model = plsdafit(X,class,comp,pret_type,assign_method,doqtlimit)
if nargin < 6
    doqtlimit = 0;
end
if iscell(class)
    class_string = class;

```

```

    [class,class_labels] = calc_class_string(class_string);
else
    class_string = {};
    class_labels = {};
end
% unfold classes
y = zeros(size(X,1),max(class));
for g=1:max(class)
    y(find(class==g),g) = 1;
end
// data scaling
[X_scal,px] = data_pretreatment(X,pret_type);
[y_scal,py] = data_pretreatment(y,'none');
% pls2
[T,P,W,U,Q,B,ssq,Ro,Rv,Lo,Lv] = mypls(X_scal,y_scal,comp);
yscal_c = T*Q';
Lo = Lo(:,comp);
yc = redo_scaling(yscal_c,py);
cumvar = ssq;
expvar(1,:) = cumvar(1,:);
for k = 2:comp; expvar(k,:) = cumvar(k,:) - cumvar(k - 1,:); end
% coefficients
b = W(:,1:comp)*inv(P(:,1:comp)'*W(:,1:comp))*Q(:,1:comp)';
% T hot
fvar = sqrt(1./(diag(T'*T)/(size(X,1) - 1)));
Thot = sum((T*diag(fvar)).^2,2);
Tcont = (T*diag(fvar)*P');
% Qres
Xmod = T*P';
Qcont = X_scal - Xmod;
for i=1:size(T,1)
    Qres(i) = Qcont(i,:)*Qcont(i,:);

```

```

end
% rmsec
for g=1:size(y,2)
    res = calc_reg_param(y(:,g),yc(:,g));
    rmsec(g) = res.RMSEC;
end
% T2 and Q limits
if doqtlimit
    mlim = pca_model(X,comp,pret_type);
    tlim = mlim.settings.tlim;
    qlim = mlim.settings.qlim;
else
    tlim = NaN;
    qlim = NaN;
end
for g=1:max(class)
    mc(g) = mean(yc(find(class==g),g));
    sc(g) = std(yc(find(class==g),g));
    mnc(g) = mean(yc(find(class~=g),g));
    snc(g) = std(yc(find(class~=g),g));
    for i=1:size(X,1)
        Pc = 1./(sqrt(2*pi)*sc(g)) * exp(-0.5*((yc(i,g) - mc(g))/sc(g)).^2);
        Pnc = 1./(sqrt(2*pi)*snc(g)) * exp(-0.5*((yc(i,g) - mnc(g))/snc(g)).^2);
        prob(i,g) = Pc/(Pc + Pnc);
    end
end
% class evaluation
[tmp,class_true] = max(y');
resthr = plsdafindthr(yc,class_true');
if strcmp(assign_method,'max')
    % assigns on the maximum calculated response
    [non,assigned_class] = max(prob');

```

```

else
    % assigns on the bayesian discrimination threshold
    assigned_class = plsdafindclass(yc,resthr.class_thr);
end
class_param = calc_class_param(assigned_class',class_true');
model.type = 'plsda';
model.yc = yc;
model.prob = prob;
model.class_calc = assigned_class';
if length(class_labels) > 0
    model.class_calc_string = calc_class_string(model.class_calc,class_labels);
end
model.class_param = class_param;
model.T = T;
model.P = P;
model.U = U;
model.Q = Q;
model.W = W;
model.b = b;
model.cumvar = cumvar;
model.expvar = expvar;
model.rmsec = rmsec;
model.H = Lo;
model.Thot = Thot;
model.Tcont = Tcont;
model.Qres = Qres;
model.Qcont = Qcont;
model.settings.pret_type = pret_type;
model.settings.px = px;
model.settings.py = py;
model.settings.y = y;
model.settings.tlim = tlim;

```

```

model.settings.qlim = qlim;
model.settings.thr_val = resthr.thr_val';
model.settings.sp = resthr.sp;
model.settings.sn = resthr.sn;
model.settings.thr = resthr.class_thr;
model.settings.assign_method = assign_method;
model.settings.raw_data = X;
model.settings.class = class;
model.settings.class_string = class_string;
model.settings.mc = mc;
model.settings.sc = sc;
model.settings.mnc = mnc;
model.settings.snc = snc;
model.cv = [];
model.labels.variable_labels = {};
model.labels.sample_labels = {};
model.labels.class_labels = class_labels;
// prediction with Partial Least Squares Discriminant Analysis (PLSDA)
function pred = plsdaPred(X,model)
W = model.W;
Q = model.Q;
P = model.P;
nF = size(model.T,2);
T = model.T;
X_scal = test_pretreatment(X,model.settings.px);
% prediction
yscal_c = 0;
for k = 1:nF
    Ttest(:,k) = X_scal*W(:,k)/(W(:,k)'*W(:,k));
    yscal_c = yscal_c + Ttest(:,k)*Q(:,k)';
    X_scal = X_scal - Ttest(:,k)*P(:,k)';
end

```

```

pred.yc = redo_scaling(yscal_c,model.settings.py);
pred.T = Ttest;
% probability
for g=1:size(pred.yc,2)
    mc(g) = model.settings.mc(g);
    sc(g) = model.settings.sc(g);
    mnc(g) = model.settings.mnc(g);
    snc(g) = model.settings.snc(g);
    for i=1:size(X,1)
        Pc = 1./((sqrt(2*pi)*sc(g)) * exp(-0.5*((pred.yc(i,g) - mc(g))/sc(g)).^2);
        Pnc = 1./((sqrt(2*pi)*snc(g)) * exp(-0.5*((pred.yc(i,g) - mnc(g))/snc(g)).^2);
        prob(i,g) = Pc/(Pc + Pnc);
    end
end
pred.prob = prob;
if strcmp(model.settings.assign_method,'max')
    [non,assigned_class] = max(prob');
else
    assigned_class = plsdafindclass(pred.yc,model.settings.thr);
end
pred.class_pred = assigned_class';
if length(model.labels.class_labels) > 0
    pred.class_pred_string = calc_class_string(pred.class_pred,model.labels.class_labels);
end
% leverages
X_scal = test_pretreatment(X,model.settings.px);
pred.H = diag(Ttest*pinv(T'*T)*Ttest');
% T hot
fvar = sqrt(1./(diag(T'*T)/(size(T,1) - 1)));
pred.Thot = sum((Ttest*diag(fvar)).^2,2);
pred.Tcont = Ttest*diag(fvar)*P';
% Qres

```



```
Xmod = Ttest*P';  
Qcont = X_scal - Xmod;  
for i=1:size(X,1)  
    pred.Qres(i) = Qcont(i,:)*Qcont(i,:);  
end  
pred.Qcont = Qcont;
```

Appendix VI

MDS: Euclidean, mahalanobis, minkowski

// Variables

// dist_type 'euclidean', 'mahalanobis', 'minkowski'

// X dataset [samples x variables]

// calculation of distances between samples of X and Xnew

function D = knn_calc_dist(X,Xnew,dist_type)

if strcmp(dist_type,'mahalanobis')

 inv_covX = pinv(cov(X));

end

for i=1:size(Xnew,1)

 if strcmp(dist_type,'euclidean')

 x_in = Xnew(i,:);

 D_squares_x = (sum(x_in'.^2))*ones(1,size(X,1));

 D_squares_w = sum(X'.^2);

 D_product = - 2*(x_in*X');

 D(i,:) = (D_squares_x + D_squares_w + D_product).^0.5;

 else

 for j=1:size(X,1)

 x = Xnew(i,:);

 y = X(j,:);

 if strcmp(dist_type,'mahalanobis')

 D(i,j) = ((x - y)*inv_covX*(x - y)')^0.5;

 elseif strcmp(dist_type,'cityblock')

 D(i,j) = sum(abs(x - y));

 elseif strcmp(dist_type,'minkowski')

 p = 2;

 D(i,j) = (sum((abs(x - y)).^p))^(1/p);

 else

 [a,bc,d,p] = calcbinary(x,y);

 if strcmp(dist_type,'sm')

```

        D(i,j)=1-((a+d)/p);
    elseif strcmp(dist_type,'rt')
        D(i,j)=1-((a+d)/(p+bc));
    elseif strcmp(dist_type,'jt')
        D(i,j)=1-(a/(a+bc));
    elseif strcmp(dist_type,'gle')
        D(i,j)=1-(2*a/(2*a+bc));
    elseif strcmp(dist_type,'ct4')
        D(i,j)=1-(log2(1+a)/log2(1+a+bc));
    elseif strcmp(dist_type,'ac')
        D(i,j)=1-((2/pi)*asin(sqrt((a+d)/p)));
    end
end
end
end
end
end

```

function [a,bc,d,p] = calcbinary(x,y)

```

p = length(x);
s = sum([x; y]);
a = length(find(s==2));
bc = length(find(s==1));
d = length(find(s==0));

```

Appendix VII

Potential functions

// Variables

```
//X                dataset [samples x variables]
// class           class vector
// type            kernel type (gaussian, triangular)
// smoot           smoothing parameter [1 x classes]
// perc            percentile to define the class boundary (95%)
// pret_type       data pretreatment
// cv_type         type of cross validation
// cv_groups       number of cv groups
// num_comp        define the number of PCs to apply Potential Functions on
// selection of optimal smoothing parameter for Potential Functions by means of cross-validation
function res = potsmootsel(X,class,type,perc,pret_type,cv_type,cv_groups,num_comp)
if nargin < 8; num_comp = NaN; end
smoot_range = [0.1:0.1:1.2];
hwait = waitbar(0,'cross validating models','CreateCancelBtn','setappdata(gcf,"canceling",1)');
setappdata(hwait,'canceling',0)
for k = 1:length(smoot_range)
    if ~ishandle(hwait)
        res.er = NaN;
        res.sensitivity = NaN;
        res.specificity = NaN;
        res.smoot_prod = NaN;
        break
    elseif getappdata(hwait,'canceling')
        res.er = NaN;
        res.sensitivity = NaN;
        res.specificity = NaN;
        res.smoot_prod = NaN;
        break
    end
end
```

```

else
    waitbar(k/length(smoot_range))
    smoot_here = ones(1,max(class))*smoot_range(k);
    out = potcv(X,class,type,smoot_here,perc,pret_type,cv_type,cv_groups,num_comp);
    res.er(k,:) = out.class_param.er_smootsel;
    res.sensitivity(k,:) = out.class_param.sn_smootsel;
    res.specificity(k,:) = out.class_param.sp_smootsel;
    res.smoot_prod(k,:) = out.smoot_prod;
end
end
if ishandle(hwait)
    delete(hwait)
end
res.settings.smoot_range = smoot_range;
res.settings.type = type;
res.settings.perc = perc;
res.settings.pret_type = pret_type;
res.settings.cv_type = cv_type;
res.settings.cv_groups = cv_groups;
res.settings.num_comp = num_comp;
end
// kernel calculation for potential functions
p = 0;
s = std(X);
if strcmp(type,'gaus')
    for i=1:size(X,1)
        n = 1;
        for j=1:size(X,2)
            d = (X(i,j) - v(j));
            r = smoot*s(j);
            n1 = 1/(r*(2*pi)^(1/2));
            n2 = -(d^2)/(2*(r^2));

```

```

        n = n*n1*exp(n2);
    end
    p = p + n/size(X,1);
end
elseif strcmp(type,'tria')
    for i=1:size(X,1)
        n = norm(X(i,:) - v);
        n = n/smoot;
        if n <= 1
            n = 1 - n;
        else
            n = 0;
        end
        p = p + n;
    end
    p = p/size(X,1);
end
end
// cross validation for class modeling Potential Functions
if iscell(class)
    class_string = class;
    [class,class_labels] = calc_class_string(class_string);
else
    class_string = {};
    class_labels = {};
end
if nargin < 9; num_comp = NaN; end
y = class;
x = X;
nobj=size(x,1);
if strcmp(cv_type,'boot')
    hwait = waitbar(0,'bootstrap validation');

```

```

out_bootstrap = zeros(nobj,1);
assigned_class = [];
binary_class = [];
class_true = [];
smoot_prod = ones(1,max(class));
for i=1:cv_groups
    waitbar(i/cv_groups)
    out = ones(nobj,1);
    whos_in = [];
    for k=1:nobj
        r = ceil(rand*nobj);
        whos_in(k) = r;
    end
    out(whos_in) = 0;
    % counters for left out samples
    out_bootstrap(find(out == 1)) = out_bootstrap(find(out == 1)) + 1;
    x_out = x(find(out == 1),:);
    x_in = x(whos_in,:);
    y_in = y(whos_in,:);
    y_out = y(find(out == 1),:);
    model = potfit(x_in,y_in,type,smoot,perc,pret_type,num_comp);
    pred = potpred(x_out,model);
    assigned_class = [assigned_class; pred.class_pred];
    binary_class = [binary_class; pred.binary_assignment];
    class_true = [class_true; class(find(out == 1))];
    for g=1:size(pred.P,2)
        smoot_prod(g) = smoot_prod(g)*prod(pred.P(find(y_out == g),g));
    end
end
class = class_true;
assigned_class = assigned_class';
delete(hwait);

```

```

elseif strcmp(cv_type,'rand')
    hwait = waitbar(0,'montecarlo validation');
    assigned_class = [];
    binary_class = [];
    smoot_prod = ones(1,max(class));
    out_rand = zeros(nobj,1);
    perc_in = 0.8;
    take_in = round(nobj*perc_in);
    class_true = [];
    for i=1:cv_groups
        waitbar(i/cv_groups)
        out = ones(nobj,1);
        whos_in = randperm(nobj);
        whos_in = whos_in(1:take_in);
        out(whos_in) = 0;
        % counters for left out samples
        out_rand(find(out == 1)) = out_rand(find(out == 1)) + 1;
        x_out = x(find(out == 1),:);
        x_in = x(whos_in,:);
        y_in = y(whos_in,:);
        y_out = y(find(out == 1),:);
        model = potfit(x_in,y_in,type,smoot,perc,pret_type,num_comp);
        pred = potpred(x_out,model);
        assigned_class = [assigned_class; pred.class_pred];
        binary_class = [binary_class; pred.binary_assignment];
        class_true = [class_true; class(find(out == 1))];
        for g=1:size(pred.P,2)
            smoot_prod(g) = smoot_prod(g)*prod(pred.P(find(y_out == g),g));
        end
    end
end
class = class_true;
assigned_class = assigned_class';

```



```

delete(hwait);
else
    obj_in_block = fix(nobj/cv_groups);
    left_over = mod(nobj,cv_groups);
    smoot_prod = ones(1,max(class));
    st = 1;
    en = obj_in_block;
    for i = 1:cv_groups
        in = ones(size(x,1),1);
        if strcmp(cv_type,'vene') % venetian blinds
            out = [i:cv_groups:nobj];
        else % contiguous blocks
            if left_over == 0
                out = [st:en];
                st = st + obj_in_block; en = en + obj_in_block;
            else
                if i < cv_groups - left_over
                    out = [st:en];
                    st = st + obj_in_block; en = en + obj_in_block;
                elseif i < cv_groups
                    out = [st:en + 1];
                    st = st + obj_in_block + 1; en = en + obj_in_block + 1;
                else
                    out = [st:nobj];
                end
            end
        end
    end
    in(out) = 0;
    x_in = x(find(in),:);
    y_in = y(find(in),:);
    x_out = x(find(in == 0),:);
    y_out = y(find(in == 0),:);

```

```

model = potfit(x_in,y_in,type,smoot,perc,pret_type,num_comp);
pred = potpred(x_out,model);
assigned_class(find(in == 0)) = pred.class_pred;
binary_class(find(in == 0),:) = pred.binary_assignment;
for g=1:size(pred.P,2)
    smoot_prod(g) = smoot_prod(g)*prod(pred.P(find(y_out == g),g));
end
end
end
cv.class_param = calc_class_param(assigned_class',class);
for g=1:size(binary_class,2)
    class_here = 2*ones(length(class),1);
    class_here(find(class == g)) = 1;
    class_here_calc = binary_class(:,g);
    class_here_calc(find(class_here_calc == 0)) = 2;
    cp_class = calc_class_param(class_here_calc,class_here);
    cv.class_param.sn_smootsel(g) = cp_class.sensitivity(1);
    cv.class_param.sp_smootsel(g) = cp_class.specificity(1);
    cv.class_param.er_smootsel(g) = cp_class.er;
end
cv.class_pred = assigned_class';
if length(class_labels) > 0
    cv.class_pred_string = calc_class_string(cv.class_pred,class_labels);
end
cv.smoot_prod = smoot_prod;
cv.settings.type = type;
cv.settings.perc = perc;
cv.settings.smoot = smoot;
cv.settings.cv_groups = cv_groups;
cv.settings.cv_type = cv_type;
cv.settings.pret_type = pret_type;
cv.settings.num_comp = num_comp;

```

```

// class identification with potential functions
function [assigned_class,binary_assignment] = potfindclass(P,thr)
for k=1:size(P,1)
    which_class = zeros(1,size(P,2));
    for g=1:size(P,2)
        if P(k,g) > thr(g)
            which_class(g) = 1;
        end
    end
    if length(find(which_class == 1)) == 1
        assigned_class(k,1) = find(which_class == 1);
    else
        assigned_class(k,1) = 0;
    end
    binary_assignment(k,:) = which_class;
end
end
// fit class modeling Potential Functions
if iscell(class)
    class_string = class;
    [class,class_labels] = calc_class_string(class_string);
else
    class_string = {};
    class_labels = {};
end
if nargin < 7; num_comp = NaN; end
if isnan(num_comp)
    [X_scal,param] = data_pretreatment(X,pret_type);
    model_pca = NaN;
else
    param = NaN;
    if num_comp == 0

```

```

    comphere = 10;
else
    comphere = num_comp;
end
model_pca = pca_model(X,comphere,pret_type);
if num_comp == 0
    compin = length(find(model_pca.E > mean(model_pca.settings.Efull)));
    model_pca = pca_model(X,compin,pret_type);
end
X_scal = model_pca.T;
end
for g=1:max(class)
    % potential function
    Xclass{g} = X_scal(find(class == g),:);
    for i=1:size(X_scal,1)
        P(i,g) = potcalc(X_scal(i,:),Xclass{g},type,smoot(g));
    end
    Pin = P(find(class == g),g);
    thr(g) = find_thr(Pin,perc);
end
class_calc = potfindclass(P,thr);
class_param = calc_class_param(class_calc,class);
model.type = 'pf';
model.P = P;
model.class_calc = class_calc;
if length(class_labels) > 0
    model.class_calc_string = calc_class_string(model.class_calc,class_labels);
end
model.class_param = class_param;
model.settings.thr = thr;
model.settings.smoot = smoot;
model.settings.type = type;

```

```

model.settings.perc = perc;
model.settings.Xclass = Xclass;
model.settings.num_comp = num_comp;
model.settings.pret_type = pret_type;
model.settings.param = param;
model.settings.model_pca = model_pca;
model.settings.raw_data = X;
model.settings.class = class;
model.cv = [];
model.labels.variable_labels = {};
model.labels.sample_labels = {};
model.labels.class_labels = class_labels;
end
function thr = find_thr(P,perc)
% class thresholds based on percentiles
q = perc*length(P)/100;
j = fix(q);
Ssort = -sort(-P);
thr = Ssort(j) + (q - j)*(Ssort(j+1) - Ssort(j));
end
// prediction with class modeling Potential Functions
smoot = model.settings.smoot;
kernel_type = model.settings.type;
if isnan(model.settings.num_comp)
    X_scal = test_pretreatment(X,model.settings.param);
else
    [model_pca] = pca_project(X,model.settings.model_pca);
    X_scal = model_pca.Tpred;
end
% calc potential
for g=1:length(model.settings.Xclass)
    for i=1:size(X_scal,1)

```

```
    P(i,g) = potcalc(X_scal(i,:),model.settings.Xclass{g},kernel_type,smoot(g));
end
end
[class_pred,binary_assignment] = potfindclass(P,model.settings.thr);
pred.P = P;
pred.class_pred = class_pred;
if length(model.labels.class_labels) > 0
    pred.class_pred_string = calc_class_string(pred.class_pred,model.labels.class_labels);
end
pred.binary_assignment = binary_assignment;
end
```

Appendix VIII

Backpropagation neural networks

//Variables

```
// X          dataset;
// class      class vector;
// settings   defined with the backpropagation settings routine;
// cv_type    type of cross validation;
// cv_groups  number of cv groups;
```

// cross-validation of Backpropagation Neural Networks (BPNN)

```
function cv = backpropagationcv(X,class,settings,cv_type,cv_groups);
```

```
if iscell(class)
```

```
    class_string = class;
```

```
    [class,class_labels] = calc_class_string(class_string);
```

```
else
```

```
    class_string = {};
```

```
    class_labels = {};
```

```
end
```

```
settings.doplot = 0;
```

```
nobj = size(X,1);
```

```
if strcmp(cv_type,'boot')
```

```
    hwait = waitbar(0,'bootstrap validation');
```

```
    out_bootstrap = zeros(nobj,1);
```

```
    class_pred = [];
```

```
    class_true = [];
```

```
    for i=1:cv_groups
```

```
        waitbar(i/cv_groups)
```

```
        out = ones(nobj,1);
```

```
        whos_in = [];
```

```
        for k=1:nobj
```

```
            r = ceil(rand*nobj);
```

```
            whos_in(k) = r;
```

```

end
out(whos_in) = 0;
% counters for left out samples
boot_how_many_out(i)=length(find(out == 1));
out_bootstrap(find(out == 1)) = out_bootstrap(find(out == 1)) + 1;
Xext = X(find(out == 1),:);
class_ext = class(find(out == 1));
Xtrain = X(whos_in,:);
class_train = class(whos_in);
model = backpropagationfit(Xtrain,class_train,settings);
pred = backpropagationpred(Xext,model);
class_pred = [class_pred; pred.class_pred];
class_true = [class_true; class_ext];
end
class = class_true;
delete(hwait);
elseif strcmp(cv_type,'rand')
hwait = waitbar(0,'montecarlo validation');
out_rand = zeros(nobj,1);
perc_in = 0.8;
take_in = round(nobj*perc_in);
class_pred = [];
class_true = [];
for i=1:cv_groups
waitbar(i/cv_groups)
out = ones(nobj,1);
whos_in = randperm(nobj);
whos_in = whos_in(1:take_in);
out(whos_in) = 0;
% counters for left out samples
out_rand(find(out == 1)) = out_rand(find(out == 1)) + 1;
Xext = X(find(out == 1),:);

```



```

class_ext = class(find(out == 1));
Xtrain = X(whos_in,:);
class_train = class(whos_in);
model = backpropagationfit(Xtrain,class_train,settings);
pred = backpropagationpred(Xext,model);
class_pred = [class_pred; pred.class_pred];
class_true = [class_true; class_ext];
end
class = class_true;
delete(hwait);
else
class_pred = zeros(size(X,1),1);
obj_in_block = fix(nobj/cv_groups);
left_over = mod(nobj,cv_groups);
st = 1;
en = obj_in_block;
for i = 1:cv_groups
in = ones(size(X,1),1);
if strcmp(cv_type,'vene') % venetian blinds
out = [i:cv_groups:nobj];
else % contiguous blocks
if left_over == 0
out = [st:en];
st = st + obj_in_block; en = en + obj_in_block;
else
if i < cv_groups - left_over
out = [st:en];
st = st + obj_in_block; en = en + obj_in_block;
elseif i < cv_groups
out = [st:en + 1];
st = st + obj_in_block + 1; en = en + obj_in_block + 1;
else

```

```

        out = [st:nobj];
    end
end
end
in(out) = 0;
Xtrain = X(find(in==1),:);
class_train = class(find(in==1));
Xext = X(find(in==0),:);
class_ext = class(find(in==0));
model = backpropagationfit(Xtrain,class_train,settings);
pred = backpropagationpred(Xext,model);
class_pred(find(in==0)) = pred.class_pred;
end
end
class_param = calc_class_param(class_pred,class);
cv.class_pred = class_pred;
if length(class_labels) > 0
    cv.class_pred_string = calc_class_string(cv.class_pred,class_labels);
end
cv.class_param = class_param;
cv.settings.cv_groups = cv_groups;
cv.settings.cv_type = cv_type;
cv.settings.backpropagation_settings = settings;
// assign samples for Backpropagation Neural Networks on the basis of thresholds and
// calculated responses
function assigned_class = backpropagationfindclass(yc,class_thr,assignment_type,yc_scal);
nobj = size(yc,1);
nclass = size(yc,2);
if strcmp(assignment_type,'thr')
    for i = 1:nobj
        pred = yc(i,:);
        chk_ass = zeros(1,nclass);

```

```

for c = 1:nclass
    if pred(c) > class_thr(c); chk_ass(c) = 1; end
end
if length(find(chk_ass)) == 1
    assigned_class(i,1) = find(chk_ass);
else
    assigned_class(i,1) = 0;
end
end
elseif strcmp(assignment_type,'max')
    [~,param] = data_pretreatment(yc_scal,'rang');
    yc_range = test_pretreatment(yc,param);
    for i = 1:nobj
        pred = yc_range(i,:);
        [a,b] = max(pred);
        assigned_class(i,1) = b;
    end
end
end
function model = backpropagationfit(X,class,settings);
if iscell(class)
    class_string = class;
    [class,class_labels] = calc_class_string(class_string);
else
    class_string = {};
    class_labels = {};
end
% backpropagation settings
scaling = 'auto';
num_hidden_neurons = settings.num_hidden_neurons;
learning_rate = settings.learning_rate;
alpha = settings.alpha;

```

```

iter_max = settings.iter;
assignment_type = settings.assignment_type;
doplot = settings.doplot;
output = zeros(size(X,1),max(class));
for g=1:max(class)
    output(find(class==g),g) = 1;
end

%STEP 1 : Normalize the Input
[X_scal,param] = data_pretreatment(X,scaling);
% add ones for bias
X_scal = [X_scal ones(size(X,1),1)];
%Find the size of Input and Output Vectors
num_var = size(X_scal,2);
num_responses = size(output,2);
%Initialize the weight matrices with random weights
init_weights = [num_var num_hidden_neurons num_responses];
for k=1:length(init_weights)-1
    W{k} = randn(init_weights(k),init_weights(k+1));
end
% initialize weights
for k=1:length(W); delta_W{k} = zeros(size(W{k})); end
iter = 0;
error_latest = 1;
%Calling function for training the neural network
if doplot; figure; set(gcf,'color','white'); end
while iter < iter_max
    iter = iter + 1;
    % Change the weight matrix W by adding delta values to them
    for k=1:length(W); W{k} = W{k} + delta_W{k}; end
    [error(iter),          delta_W,          ~,output_pred_tmp] =
nettrain(X_scal,output,W,learning_rate,alpha,delta_W);

```

```

error_latest = error(iter);
if doplot > 0
    thr_tmp = findthr(output_pred_tmp{end},class,1);
    class_calc_tmp
backpropagationfindclass(output_pred_tmp{end},thr_tmp.class_thr,assignment_type,output_pre
d_tmp{end});
    class_param_tmp = calc_class_param(class_calc_tmp,class);
    ner(iter) = class_param_tmp.ner;
    sensitivity(iter,:) = class_param_tmp.sensitivity;
    if doplot > 1
        updateplot(error(1:iter),1,iter_max,ner,sensitivity);
    end
end
end
output_pred = backpropagationproject(X_scal,W);
res = findthr(output_pred{end},class);
class_calc
backpropagationfindclass(output_pred{end},res.class_thr,assignment_type,output_pred{end});
class_param = calc_class_param(class_calc,class);
if doplot > 0; updateplot(error(1:iter),1,iter_max,ner,sensitivity);end
model.type = 'backprop';
model.W = W;
model.output_pred = output_pred{end};
model.class_calc = class_calc;
if length(class_labels) > 0
    model.class_calc_string = calc_class_string(model.class_calc,class_labels);
end
model.class_param = class_param;
model.settings.network_settings = settings;
model.settings.param = param;
model.settings.thr = res.class_thr;
model.settings.thr_val = res.thr_val';

```

```

model.settings.sp = res.sp;
model.settings.sn = res.sn;
model.settings.error = error;
model.settings.raw_data = X;
model.settings.class = class;
model.cv = [];
model.labels.variable_labels = {};
model.labels.sample_labels = {};
model.labels.class_labels = class_labels;
end

function [E, delta_W, residuals, Output_of_HiddenLayer]
= nettrain(X,Output,W,learning_rate,alpha,delta_W)
%Calculating the Output of Input Layer
%Output of Input Layer is same as the Input of Input Layer
[Output_of_HiddenLayer,Input_of_HiddenLayer] = backpropagationproject(X,W);
%Now we need to calculate the Error using Root Mean Square method
[E, residuals] = calc_errors(Output,Output_of_HiddenLayer{end});
%Calculate the matrix 'd' with respect to the desired output
d = (Output - Output_of_HiddenLayer{end});
d = d.*Output_of_HiddenLayer{end};
d = d.*(1-Output_of_HiddenLayer{end});
%Calculating delta output layer
delta_W{end} = alpha*delta_W{end} + learning_rate.*(d.*Output_of_HiddenLayer{end-1});
%Calculating error matrix
error = (W{end}*d)';
for k = 1:length(W) - 2
    %Calculating d*
    d_star = [];
    d_star = error.*Output_of_HiddenLayer{end-k};
    d_star = d_star.*(1-Output_of_HiddenLayer{end-k});
    %Calculating delta W

```

```

    delta_W{end-k} = alpha*delta_W{end-k} + learning_rate*Input_of_HiddenLayer{end-k-
1}'*d_star;
    % updating error for next correction
    error = (W{end-k}*d_star)';
end
d_star = [];
d_star = error.*Output_of_HiddenLayer{1};
d_star = d_star.*(1-Output_of_HiddenLayer{1});
% Calculating delta W
s1 = alpha*delta_W{1};
s2 = X'*d_star;
delta_W{1} = s1+learning_rate*s2;
end
%-----
function [E, residuals] = calc_errors(Output,Output_of_HiddenLayer)
difference = Output - Output_of_HiddenLayer;
square = difference.*difference;
E = sum(square(:))/size(Output,1);
residuals = sum(square,2);
end
%-----
function res = findthr(yc,class,dofast)
if nargin < 3
    dofast = 0;
end
if dofast == 0
    % calc values for AUC
    rsize = 100;
    for g=1:size(yc,2)
        class_in = ones(size(class,1),1);
        class_in(find(class ~= g)) = 2;
        count = 0;

```

```

y_in = yc(:,g);
miny = min(y_in) - min(y_in)/10;
thr = max(y_in) + max(y_in)/10;
step = (thr - miny)/rsize;
if (thr - miny) > 0.01
    while thr > miny
        count = count + 1;
        class_calc_in = ones(size(class,1),1);
        thr = thr - step;
        sample_out_g = find(y_in < thr);
        class_calc_in(sample_out_g) = 2;
        cp = calc_class_param(class_calc_in,class_in);
        sp(count,g) = cp.specificity(1);
        sn(count,g) = cp.sensitivity(1);
        thr_val(count,g) = thr;
    end
else
    sp = NaN;
    sn = NaN;
    thr_val = NaN;
end
end
res.sp = sp;
res.sn = sn;
res.thr_val = thr_val;
end
% find best thr based on bayesian discrimination threshold
for g=1:max(class)
    P_g = yc(find(class==g),g);
    P_notg = yc(find(class~=g),g);
    m_g = mean(P_g); s_g = std(P_g);
    m_notg = mean(P_notg); s_notg = std(P_notg);

```



```

stp = abs(m_g - m_notg)/1000;
where = [m_notg:stp:m_g];
% fit normal distribution
% npdf_g = normpdf(where,m_g,s_g);
x_g = (where - m_g) ./ s_g;
npdf_g = exp(-0.5 * x_g.^2) ./ (sqrt(2*pi) .* s_g);
%npdf_notg = normpdf(where,m_notg,s_notg);
x_notg = (where - m_notg) ./ s_notg;
npdf_notg = exp(-0.5 * x_notg.^2) ./ (sqrt(2*pi) .* s_notg);
minval = NaN;
for k=1:length(where)
    diff = abs(npdf_g(k)-npdf_notg(k));
    if isnan(minval)|diff < minval
        minval = diff;
        class_thr(g) = where(k);
    end
end
if isnan(minval)
    class_thr(g) = mean([m_g m_notg]);
end
end
res.class_thr = class_thr;
end
%-----
function updateplot(error,start_epoch,end_epoch,ner,sensitivity)
% plot residuals
subplot(2,1,1)
drawnow
cla
hold on
plot(error,'LineWidth',1.5,'Color',c1);
plot(length(error),error(end),'o','MarkerEdgeColor',c1,'MarkerSize',4,'MarkerFaceColor',c1);

```

```

% grid and axis
grid on
ax = gca;
ax.GridLineStyle = ':';
ax.GridAlpha = 0.5;
M = max(error); M = M*1.1;
axis([start_epoch end_epoch 0 M])
ylabel('error')
xlabel('epochs')
title(['training epochs: ' num2str(length(error)) ' of ' num2str(end_epoch)])
hold off
box on
% plot ner
subplot(2,1,2)
drawnow
cla
hold on
c2 = [0.8500 0.3250 0.0980];
c3 = [1 0.6 0.4];
for g=1:size(sensitivity,2)
    plot(sensitivity(:,g),'LineWidth',1,'Color',c3);
    plot(length(ner),sensitivity(end,g),'o','MarkerEdgeColor',c3,'MarkerSize',2.5,'MarkerFaceColor',c
3);
end
plot(ner,'LineWidth',1.5,'Color',c2);
plot(length(ner),ner(end),'o','MarkerEdgeColor',c2,'MarkerSize',4,'MarkerFaceColor',c2);
grid on
ax = gca;
ax.GridLineStyle = ':';
ax.GridAlpha = 0.5;
m = min(ner); m = m*0.9;
axis([start_epoch end_epoch m 1])

```

```

ylabel('NER and sensitivities')
xlabel('epochs')
hold off
box on
end

// prediction with Backpropagation Neural Networks (BPNN)
function pred = backpropagationpred(X,model);
%STEP 1 : Normalize the Input
[X_scal] = test_pretreatment(X,model.settings.param);
X_scal = [X_scal ones(size(X,1),1)];
output_pred = backpropagationproject(X_scal,model.W);
class_pred =
backpropagationfindclass(output_pred{end},model.settings.thr,model.settings.network_settings.
assignment_type,model.output_pred);
pred.output_pred = output_pred{end};
pred.class_pred = class_pred;
if length(model.labels.class_labels) > 0
    pred.class_pred_string = calc_class_string(pred.class_pred,model.labels.class_labels);
end
end

// define network settings for Backpropagation Neural Networks (BPNN)
function settings = backpropagationsettings(num_hidden_neurons,learning_rate)
settings.num_hidden_neurons = num_hidden_neurons;
settings.learning_rate = learning_rate;
settings.alpha = 0.5;
settings.iter = 1000;
settings.assignment_type = 'thr';
settings.doplot = 1;
end

function class_param = calc_class_param(class_calc,class)
num_class = max([max(class) max(class_calc)]);
nobj = size(class,1);

```

```

conf_mat = zeros(num_class,num_class+1);
for g = 1:num_class
    in_class = find(class==g);
    for k = 1:num_class
        conf_mat(g,k) = length(find(class_calc(in_class) == k));
    end
    conf_mat(g,num_class + 1) = length(find(class_calc(in_class) == 0));
end
// sensitivity, specificity, precision, class error, accuracy
accuracy = 0;
for g = 1:num_class
    if sum(conf_mat(:,g)) > 0
        precision(g) = conf_mat(g,g)/sum(conf_mat(:,g));
        sensitivity(g) = conf_mat(g,g)/sum(conf_mat(g,1:num_class));
    else
        precision(g) = 0;
        sensitivity(g) = 0;
    end
    in = ones(num_class,1); in(g) = 0;
    red_mat = conf_mat(find(in),1:num_class);
    specificity(g) = 0;
    for k = 1:size(red_mat,2)
        if k ~= g; specificity(g) = specificity(g) + sum(red_mat(:,k)); end;
    end
    if sum(sum(red_mat)) > 0
        specificity(g) = specificity(g)/sum(sum(red_mat));
    else
        specificity(g) = 0;
    end
    accuracy = accuracy + conf_mat(g,g);
end
accuracy = accuracy/sum(sum(conf_mat(:,1:num_class)));

```

```
% error rates
not_ass = sum(conf_mat(:,end))/nobj;
ner = mean(sensitivity);
er = 1 - ner;
class_param.conf_mat = conf_mat;
class_param.ner = ner;
class_param.er = er;
class_param.accuracy = accuracy;
class_param.not_ass = not_ass;
class_param.precision = precision;
class_param.sensitivity = sensitivity;
class_param.specificity = specificity;
```

Appendix IX

Support vector machine (SVM)

// Variables

```
// X                dataset [samples x variables]
// class            class vector
// kernel           type of kernel (linear, polynomial, rbf)
// pret_type        data pretreatment
// cv_type          type of cross validation
// cv_groups        number of cv groups
// num_comp         define the number of PCs to apply SVM (integer_number)
// C                upper bound for the coefficients alpha during training (cost)
// kernelpar        parameter for rbf and poly kernels
// model            SVM model calculated by means of svmfit

// selection of the optimal C (cost, upper bound for the coefficients alpha) and kernal param (only
// for rbf and poly kernels) for Support Vector Machines by means of cross validation
function res = svmcostsel(X,class,kernel,pret_type,cv_type,cv_groups,num_comp)
if nargin < 7; num_comp = NaN; end
C_seq = [0.1 1 10 100 1000];
if strcmp('linear',kernel)
    kernalparam_seq = [];
else
    kernalparam_seq = [0.05 0.07 0.10 0.14 0.20 0.28 0.40 0.57 0.80 1.13 1.60
                      2.26 3.20 4.53 6.40 9.00];
end
hwait = waitbar(0,'cross validating models','CreateCancelBtn','setappdata(gcf,'canceling',1)');
setappdata(hwait,'canceling',0)
for c = 1:length(C_seq)
    if ~ishandle(hwait)
        res.er = NaN;
        res.ner = NaN;
        res.average_svind = NaN;
    end
end
```

```

    break
elseif getappdata(hwait,'canceling')
    res.er = NaN;
    res.ner = NaN;
    res.average_svind = NaN;
    break
else
    waitbar(c/length(C_seq))
    C = C_seq(c);
    if strcmp('linear',kernel)
        kernelpar = [];
        out = svmcv(X,class,kernel,C,kernelpar,pret_type,cv_type,cv_groups,num_comp);
        res.er(c) = out.class_param.er;
        res.ner(c) = out.class_param.ner;
        res.average_svind(c) = out.average_svind;
    else
        for k = 1:length(kernalparam_seq)
            kernelpar = kernalparam_seq(k);
            % disp(['cross validating C: ' num2str(C) ' and param: ' num2str(kernelpar)])
            out = svmcv(X,class,kernel,C,kernelpar,pret_type,cv_type,cv_groups,num_comp);
            res.er(c,k) = out.class_param.er;
            res.ner(c,k) = out.class_param.ner;
            res.average_svind(c,k) = out.average_svind;
        end
    end
end
end
end
if ishandle(hwait)
    delete(hwait)
end
res.kernalparam_seq = kernalparam_seq;
res.cost_seq = C_seq;

```

```

res.settings.kernel = kernel;
res.settings.cv_type = cv_type;
res.settings.cv_groups = cv_groups;
res.settings.num_comp = num_comp;
res.settings.pret_type = pret_type;
// cross-validation for Support Vector Machines (only two classes allowed)
function cv = svmcv(X,class,kernel,C,kernelpar,pret_type,cv_type,cv_groups,num_comp)
if iscell(class)
    class_string = class;
    [class,class_labels] = calc_class_string(class_string);
else
    class_string = {};
    class_labels = {};
end
if nargin < 9; num_comp = NaN; end
nobj = size(X,1);
if strcmp(cv_type,'boot')
    hwait = waitbar(0,'bootstrap validation');
    svind = zeros(cv_groups,1);
    out_bootstrap = zeros(nobj,1);
    class_pred = [];
    class_true = [];
    for i=1:cv_groups
        waitbar(i/cv_groups)
        out = ones(nobj,1);
        whos_in = [];
        for k=1:nobj
            r = ceil(rand*nobj);
            whos_in(k) = r;
        end
        out(whos_in) = 0;
        % counters for left out samples

```



```

boot_how_many_out(i)=length(find(out == 1));
out_bootstrap(find(out == 1)) = out_bootstrap(find(out == 1)) + 1;
Xext = X(find(out == 1),:);
class_ext = class(find(out == 1));
Xtrain = X(whos_in,:);
class_train = class(whos_in);
model = svmfit(Xtrain,class_train,kernel,C,kernelpar,pret_type,num_comp);
pred = svmpred(Xext,model);
class_pred = [class_pred; pred.class_pred];
class_true = [class_true; class_ext];
svind(i) = length(model.svind);
end
class = class_true;
delete(hwait);
elseif strcmp(cv_type,'rand')
hwait = waitbar(0,'montecarlo validation');
svind = zeros(cv_groups,1);
out_rand = zeros(nobj,1);
perc_in = 0.8;
take_in = round(nobj*perc_in);
class_pred = [];
class_true = [];
for i=1:cv_groups
waitbar(i/cv_groups)
out = ones(nobj,1);
whos_in = randperm(nobj);
whos_in = whos_in(1:take_in);
out(whos_in) = 0;
% counters for left out samples
out_rand(find(out == 1)) = out_rand(find(out == 1)) + 1;
Xext = X(find(out == 1),:);
class_ext = class(find(out == 1));

```

```

Xtrain = X(whos_in,:);
class_train = class(whos_in);
model = svmfit(Xtrain,class_train,kernel,C,kernelpar,pret_type,num_comp);
pred = svmpred(Xext,model);
class_pred = [class_pred; pred.class_pred];
class_true = [class_true; class_ext];
svind(i) = length(model.svind);
end
class = class_true;
delete(hwait);
else
svind = zeros(cv_groups,1);
class_pred = zeros(size(X,1),1);
obj_in_block = fix(nobj/cv_groups);
left_over = mod(nobj,cv_groups);
st = 1;
en = obj_in_block;
for i = 1:cv_groups
in = ones(size(X,1),1);
if strcmp(cv_type,'vene') % venetian blinds
out = [i:cv_groups:nobj];
else % contiguous blocks
if left_over == 0
out = [st:en];
st = st + obj_in_block; en = en + obj_in_block;
else
if i < cv_groups - left_over
out = [st:en];
st = st + obj_in_block; en = en + obj_in_block;
elseif i < cv_groups
out = [st:en + 1];
st = st + obj_in_block + 1; en = en + obj_in_block + 1;

```

```

        else
            out = [st:nobj];
        end
    end
end
end
in(out) = 0;
Xtrain = X(find(in==1),:);
class_train = class(find(in==1));
Xext = X(find(in==0),:);
model = svmfit(Xtrain,class_train,kernel,C,kernelpar,pret_type,num_comp);
pred = svmpred(Xext,model);
class_pred(find(in==0)) = pred.class_pred;
svind(i) = length(model.svind);
end
end
class_param = calc_class_param(class_pred,class);
cv.class_pred = class_pred;
if length(class_labels) > 0
    cv.class_pred_string = calc_class_string(cv.class_pred,class_labels);
end
cv.average_svind = mean(svind);
cv.class_param = class_param;
cv.settings.cv_groups = cv_groups;
cv.settings.cv_type = cv_type;
cv.settings.kernel = kernel;
cv.settings.C = C;
cv.settings.kernelpar = kernelpar;
cv.settings.pret_type = pret_type;
cv.settings.num_comp = num_comp;
// fit Support Vector Machines (only two classes allowed)
function model = svmfit(X,class,kernel,C,kernelpar,pret_type,num_comp)
if iscell(class)

```

```

class_string = class;
[class,class_labels] = calc_class_string(class_string);
else
class_string = {};
class_labels = {};
end
if nargin < 7; num_comp = NaN; end
if max(class) > 2
disp('more than two classes detected, but only two classes allowed! model wont be calculated')
model = NaN;
return;
end
class(find(class == 2)) = -1;
tol = 1e-2;
% pretreat data
if isnan(num_comp)
[X_scal,param] = data_pretreatment(X,pret_type);
model_pca = NaN;
else
param = NaN;
if num_comp == 0
comphere = 10;
else
comphere = num_comp;
end
model_pca = pca_model(X,comphere,pret_type);
if num_comp == 0
compin = length(find(model_pca.E > mean(model_pca.settings.Efull)));
model_pca = pca_model(X,compin,pret_type);
end
X_scal = model_pca.T;
end

```

```

net = fitcsvm(X_scal,class,'KernelFunction',kernel,'KernelScale',kernelpar,'BoxConstraint',C);
[~,dist] = predict(net,X_scal);
dist = dist(:,2);
net_scores = fitPosterior(net,X_scal,class);
[~,prob] = predict(net_scores,X_scal);
prob = prob(:,[2 1]);
alpha = zeros(size(X,1),1);
alpha(find(net.IsSupportVector)) = net.Alpha;
% class prediction
class_calc = sign(dist);
class(find(class == -1)) = 2;
class_calc(find(class_calc == -1)) = 2;
class_param = calc_class_param(class_calc,class);
% store linear coefficients and bias
model.type = 'svm';
model.alpha = alpha;
model.svind = find(net.IsSupportVector);
model.b = net.Beta;
model.bias = net.Bias;
model.dist = dist;
model.prob = prob;
model.class_calc = class_calc;
if length(class_labels) > 0
    model.class_calc_string = calc_class_string(model.class_calc,class_labels);
end
model.class_param = class_param;
model.settings.net = net;
model.settings.net_scores = net_scores;
model.settings.param = param;
model.settings.pret_type = pret_type;
model.settings.C = C;
model.settings.kernel = kernel;

```

```

model.settings.kernelpar = kernelpar;
model.settings.svind_data_scaled = X_scal(model.svind,:);
model.settings.svind_data = X(model.svind,:);
model.settings.data_scaled = X_scal;
model.settings.num_comp = num_comp;
model.settings.model_pca = model_pca;
model.settings.raw_data = X;
model.settings.class = class;
model.cv = [];
model.labels.variable_labels = {};
model.labels.sample_labels = {};
model.labels.class_labels = class_labels;
// prediction with Support Vector Machines
function pred = svmpred(X,model)
if isnan(model.settings.num_comp)
    X_scal = test_pretreatment(X,model.settings.param);
else
    [model_pca] = pca_project(X,model.settings.model_pca);
    X_scal = model_pca.Tpred;
end
[~,dist] = predict(model.settings.net,X_scal);
dist = dist(:,2);
[~,prob] = predict(model.settings.net_scores,X_scal);
prob = prob(:,[2 1]);
% class prediction
class_pred = sign(dist);
% put 2 instead of -1
class_pred(find(class_pred == -1)) = 2;
pred.class_pred = class_pred;
if length(model.labels.class_labels) > 0
    pred.class_pred_string = calc_class_string(pred.class_pred,model.labels.class_labels);
end

```

```

pred.prob = prob;
pred.dist = dist;
// pretreatment for test data
// X:    data matrix [samples x variables]
// param:  output data structure from data_pretreatment routine
function [X_scal] = test_pretreatment(X,param)
a = param.a;
s = param.s;
m = param.m;
M = param.M;
pret_type = param.pret_type;
if strcmp(pret_type,'cent')
    amat = repmat(a,size(X,1),1);
    X_scal = X - amat;
elseif strcmp(pret_type,'scal')
    f = find(s>0);
    smat = repmat(s,size(X,1),1);
    X_scal = zeros(size(X,1),size(X,2));
    X_scal = X(:,f)./smat(:,f);
elseif strcmp(pret_type,'auto')
    f = find(s>0);
    amat = repmat(a,size(X,1),1);
    smat = repmat(s,size(X,1),1);
    X_scal = zeros(size(X,1),size(X,2));
    X_scal(:,f) = (X(:,f) - amat(:,f))./smat(:,f);
elseif strcmp(pret_type,'rang')
    f = find(M - m > 0);
    mmat = repmat(m,size(X,1),1);
    Mmat = repmat(M,size(X,1),1);
    X_scal = zeros(size(X,1),size(X,2));
    X_scal(:,f) = (X(:,f) - mmat(:,f))./(Mmat(:,f) - mmat(:,f));
else

```

```
X_scal = X;  
end
```