SPECTRAL ANALYSIS OF CANCER BIOMARKERS IN HUMAN SERUM USING A CUSTOM PORTABLE ANALYZER

by

Kaan Oktay

B.S., Electrical and Electronics Engineering, Boğaziçi University, 2016

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Electrical and Electronics Engineering
Boğaziçi University
2019

SPECTRAL ANALYSIS OF CANCER BIOMARKERS IN HUMAN SERUM USING A CUSTOM PORTABLE ANALYZER

APPROVED BY:

Prof. Burak Acar ..... (Thesis Supervisor)

Assoc. Prof. Arzucan Özgür .....

Asst. Prof. Onur Ferhanoğlu .....

DATE OF APPROVAL: 24.06.2019

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deep gratitude to my supervisor Prof. Burak Acar, for his patient guidance and thought-provoking ideas during this study. Without him, I would never find enough motivation to finish this work. I would also like to express my deep appreciation to my past supervisor Prof. Hamdi Torun, who introduced this research topic to me, for his valuable contributions not just for this study but also for my academic future. My grateful thanks are also extended to Prof. Zeynep Madak-Erdogan, for her help in all biologically-related parts and hospitality during my visit to her laboratory.

I would also like to thank to Türkiye Bilimsel ve Teknolojik Araştırma Kurumu (TÜBİTAK) for supporting me throughout my master study with 2210-A National Scholarship Programme for master students. I would like to extend my thanks to Bogazici University for supporting me and our project by research funds grant.

I would also like to thank Eylem Kulkoyluoglu Cotul and Ashlie Santaliz-Casiano for their help throughout my experiments in their laboratory and biological measurements. My thanks also extended to Meera Patel, Natascia Marino and Anna Maria V. Storniolo for conducting blood serum collection and molecule measurement processes.

I would also like to thank to Onur Ates for sharing his deep experience with material development process in clean room. I also wish to thank Alper Bayram for helping me in the coding parts of this study and for being such a good company in the laboratory.

I would also like to extend my thanks to my dear family for their support throughout my study and for their encouragement to my lifelong passion to be a scientist. And last but not the least, I would like to thank my dear girlfriend, Cagla Özsoy, for always being there for me whenever I need help or encouragement to finish this study.

ABSTRACT

SPECTRAL ANALYSIS OF CANCER BIOMARKERS IN HUMAN SERUM USING A CUSTOM PORTABLE ANALYZER

In this study, we systematically evaluate metabolites and proteins in blood to develop a pipeline to identify potential biomarkers for breast cancer risk. Our aim is to identify a group of molecules, which can be targeted in the design of portable and low-cost blood biomarker detection devices. We obtained plasma samples from women who were cancer free and women who were cancer free at the time of blood collection but developed breast cancer later. Potential prognostic biomarkers for breast cancer risk were extracted from plasma metabolomics and proteomics data using statistical and discriminative power analyses. One of these biomarkers was validated by independent biological measurements. These biomarkers can be used to develop low-cost screening methods towards early diagnosis and hence decreased mortality due to breast cancer. As a low-cost biomarker screening solution, we proposed a Fabry-Perot (FP) interferometer-based spectroscopic biomarker analysis setup to quantitatively detect specific molecules in their environments. We used commercial Fabry-Perot interferometers and their evaluation board, which takes the output of the interferometers and process it to make the measurements readable and interpretable. For the surface on which we drop our samples for the measurements, we used Calcium Fluoride (CaF_2) windows. In addition to hardware parts, we used a reconstruction method on computer generated spectra to overcome the inherited blurring problem of our interferometers.

ÖZET

İNSAN KAN SERUMUNDAKİ KANSER BİYOBELİRTEÇLERİNİN ÖZEL TASARLANMIŞ TAŞINABİLİR TAHLİL CİHAZI İLE SPEKTRAL ANALİZİ

Bu çalışmada, meme kanserine yönelik riskleri ortaya çıkarabilecek potansiyel biyobelirteçleri belirlemek amacıyla bir yöntem geliştirmek için kan içerisindeki metabolitler ve proteinler sistematik olarak değerlendirilmiştir. Bu çalışmanın amacı, taşınabilir ve düşük maliyetli, kandaki biyobelirteçleri tespit edebilecek bir cihazın dizaynında hedeflenebilecek bir molekül grubunu tespit edebilmektir. Bunun için sağlıklı kadınlardan ve kanları toplandıktan bir süre sonra kansere yakalanmış ancak kan toplanma esnasında sağlıklı olan kadınlardan kan plazması örnekleri toplanmıştır. Meme kanseri riskini önceden işaret edebilecek potansiyel biyobelirteçler, istatistiksel ve ayırt edici analizler kullanılarak kan plazmasında ölçülen metabolomik ve proteomik verisinden tespit edilmiştir. Bu biyobelirteçlerden bir tanesi bağımsız biyolojik ölçümler aracılığıyla doğrulanmıştır. Bu biyobelirteçler, meme kanserinin erken teşhisi ve dolayısıyla meme kanseri sebepli ölüm oranının düşürülmesi amacıyla düşük bütçeli takip ve ölçüm cihazları geliştirmek için kullanılabilir. Önemli moleküllerin bulundukları ortamlardaki nicel miktarını tespit etmek amacıyla, düşük bütçeli bir biyobelirteç gözlemleme ve ölçüm aracı çözümü olarak, Fabry-Perot (FP) interferometresi tabanlı bir spektroskopik biyobelirteç analizi düzeneği önerilmiş ve yapılmıştır. Bu ölçüm düzeneği için, piyasada bulunabilen FP interferometreleri ve interferometrelerin çıkış sinyalini alıp yorumlanabilir hale getiren devreler kullanılmıştır. Yüzeyine ölçüm amaçlı örneklerin damlatılması için Kalsiyum Florür (CaF_2) pencereleri kullanılmıştır. Bu donanımsal parçalara ek olarak, interferometrelerin doğasından kaynaklanan bulanıklaştırma problemini çözmek için bilgisayarda üretilmiş spektrumlar üzerinde sinyal yeniden yapılandırma yöntemleri kullanılmıştır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii

ABSTRACT iv

ÖZET v

LIST OF FIGURES viii

LIST OF TABLES x

LIST OF SYMBOLS xi

LIST OF ACRONYMS/ABBREVIATIONS xiii

1. INTRODUCTION 1

2. BACKGROUND 5

2.1. Screening Methods for Early Detection of Breast Cancer 5

2.1.1. Mammography 5

2.1.2. Ultrasound 5

2.1.3. Magnetic Resonance Imaging (MRI) 6

2.2. Biomarker Based Methods for Early Detection of Breast Cancer 6

2.3. Measurement Tools for Biomarker Based Methods for Early Detection of Breast Cancer 8

3. BLOOD PLASMA DATA 9

3.1. Patients and Plasma Samples 9

3.2. Plasma Preparation 10

3.3. OLINK Protein Biomarker and Whole Metabolite Profiling Assays . . 10

3.4. Stem Cell Factor (SCF) Quantification Using Enzyme-Linked Immunosorbent Assay (ELISA) 11

4. STATISTICAL AND BIOLOGICAL ANALYSIS 12

4.1. Preprocessing 12
4.2. Molecule Ranking and Elimination 13
4.2.1. Molecule Ranking by Student's t-Test 13
4.2.2. Molecule Elimination by Logistic Regression 14
4.2.3. Molecule Elimination by Correlation and Paired t-Test Analysis 17
4.3. Classification on Selected Molecules 18
4.4. Biological Relevance of Selected Molecules 22
4.5. Discussion of Statistical and Biological Analysis 25

5. SPECTROSCOPIC DEVICE SETUP AND SIGNAL RECONSTRUCTION 26
5.1. Measurement Setup 26
5.2. Material Development on IR-Transparent Window 29
5.3. Low Resolution Measurements by Spectroscopic Device 30
5.4. Spectral Reconstruction by Least Absolute Shrinkage and Selection Operator (LASSO): Simulation Experiments 30
5.4.1. Spectral Computer Simulations 30
5.4.2. Reconstruction by LASSO 32
5.5. Discussion of Spectral Measurements 33

6. CONCLUSIONS AND FUTURE STUDIES 36

REFERENCES 37

LIST OF FIGURES

Figure 4.1. (a) Typical ROC curve. At each point on this ROC curve, we have different thresholds ϕ and therefore get different tp- and fp-rates. (b) Different classifiers produce different ROC curves. We prefer the highest AUC. [35] 17

Figure 4.2. (A) Results on top-ranking 20 molecules. (B) Results on the molecules selected by LR classification performance. (C) Results on the molecules selected by correlation analysis. (D) Results on the molecules selected by paired t-test analysis. 23

Figure 4.3. Changes in the levels of SCF in healthy and susceptible individuals in the original data obtained by OLINK analysis. There were two probes for SCF in the original analysis. 24

Figure 4.4. Changes in the levels of SCF in healthy and susceptible individuals in additional samples. Results from Figure 4.3 are independently validated using ELISA assay. 24

Figure 5.1. Tuning mechanism of FP interferometers sold by InfraTec GmbH [39]. 27

Figure 5.2. Fabry-Perot interferometer and its closed-circuit control board which are both commercially available and sold by InfraTec GmbH. 27

Figure 5.3. Evaluation board for Fabry-Perot interferometers and controller for IR source power. 28

Figure 5.4. Transmission spectra of CaF_2 windows bought from Thorlabs [40]. 28

Figure 5.5. Overall measurement setup which includes FP interferometer, IR source, CaF_2 window and evaluation board connected to the interferometer and IR source. 29

Figure 5.6. Measured FTIR spectrum (ground truth) and FP spectrum from a random chemical sample. 30

Figure 5.7. Some example mixtures of generated FTIR sample spectra. $\mathbf{M}_i$ represent mixture of two different samples $\mathbf{F}_1$ and $\mathbf{F}_2$ in different proportions. 31

Figure 5.8. Generated ground truth FTIR spectra and blurred FP spectra for two different samples $\mathbf{F}_1$ and $\mathbf{F}_2$. 32

Figure 5.9. Reconstructed spectra of some example mixtures with their FP and FTIR spectra. 34

Figure 5.10. Comparison of peak values at different wavelengths in reconstructed and true FTIR spectra. 35

LIST OF TABLES

Table 4.1. First 20 molecules ranked by their p-values. 14

Table 4.2. Demonstration of tp, fp, tn and fn. 16

Table 4.3. Paired t-test p-values between each pair of 20 most important molecules selected by their p-values. 19

Table 4.4. Pearson's correlation coefficient between each pair of twenty most important molecules ranked by their p-values. 20

Table 4.5. Top ranking 20 molecules according to their p-values and the remaining molecules after further elimination by each method: LR classification performance, correlation and paired t-test analysis. 21

LIST OF SYMBOLS

c	Constant for soft thresholding
C_0	Positive class
C_1	Negative class
d	Distance between the reflectors of Fabry-Perot interferometers
D	Number of input features
$\mathbf{F}_i$	Spectrum of computer generated sample i
$\mathbf{H}$	Blurring (convolution) matrix
$\mathbf{M}_i$	Spectrum of mixture i of computer generated samples
n	Refractive index of the medium between the reflectors of Fabry-Perot interferometers
N	Number of all individuals
$N_{healthy}$	Number of healthy individuals
$S_{\lambda/c}(\cdot)$	Soft thresholding operator with the threshold λ/c
$\mathbf{v}$	Observed Fabry-Perot spectrum
w_i	Weight of molecule i
$x_i^{(k)}$	Measurement of molecule i for individual k
$\hat{x}_i^{(k)}$	Normalized measurement of molecule i for individual k
$y^{(k)}$	True label of individual k
$\hat{y}^{(k)}$	Predicted probability assignment to the label of individual k
$\mathbf{z}$	Underlying ground truth spectrum
β	Bias term
$\boldsymbol{\theta}$	Optimum parameters
λ	Sparsity parameter
λ_m	Detection wavelength of a Fabry-Perot interferometer for the order of interference m
$\lambda_{max}(\cdot)$	Maximum eigenvalue of the given term
$\mu_{i_{healthy}}$	Mean measurement of molecule i for healthy individuals

$\sigma_{i_{healthy}}$	Standard deviation of measurement of molecule i for healthy individuals
$\sigma(.)$	Sigmoid function
ϕ	Probability threshold for determining predicted class

LIST OF ACRONYMS/ABBREVIATIONS

ADC	Analog to Digital Converter
AMDIS	Automated Mass Spectral Deconvolution and Identification System
APOC1	Apolipoprotein C-I
AUC	Area under curve
BMI	Body mass index
BRCA1	Breast cancer type 1 protein
BRCA2	Breast cancer type 2 protein
CA	Cancer antigen
CCL5	Chemokine ligand 5
CI	Confidence interval
CTC	Circulating tumor cell
ctDNA	Circulating tumor DNA
DNA	Deoxyribonucleic acid
ELISA	Enzyme-linked immunosorbent assay
ERα	Estrogen receptor alpha
FFPE	Formalin-fixed paraffin-embedded
fn	False negatives
fp	False positives
FP	Fabry-Perot
FTIR	Fourier transform infrared spectroscopy
GC	Gas chromatography
GmbH	Gesellschaft mit beschränkter Haftung
HER2	Human epidermal growth factor receptor 2
HRP	Horseradish peroxidase
HSP90A	Heat shock protein 90A
IDC	Invasive ductal carcinoma
IGFBP3	Insulin-like growth factor-binding protein 3
IR	Infrared

IRB	Institutional review board
LR	Logistic regression
MRI	Magnetic Resonance Imaging
MS	Mass spectrometry
OPN	Osteopontin
PAI-1	Plasminogen activator inhibitor-1
PAPPA	Pappalysin-1
pCC	Pearson's correlation coefficient
PR	Progesterone receptors
RCF	Relative centrifugal force
RNA	Ribonucleic acid
ROC	Receiver Operating Characteristic
SCF	Stem cell factor
SLPI	Secretory leukocyte proteaseinhibitor
TMB	Tetramethylbenzidine
tn	True negatives
TOP2A	Topoisomerase II alpha
tp	True positives
UIUC	University of Illinois Urbana-Champaign
UKCTOCS	United Kingdom Collaborative Trial of Ovarian Cancer Screening
UV	Ultraviolet
WHO	World Health Organization

1. INTRODUCTION

Death due to breast cancer is second leading cause of death among women [1] and potentially preventable with early detection and treatment. On the other hand, according to World Health Organization (WHO), there is a sharp rise in overall number of breast cancer incidences [2], and it is the second leading cause of death among adult women world-wide due to changes in life style [3], reproductive factors [4] and increased life expectancy [5]. The problem is more significant in middle-and low-income countries, as 58% of all deaths due to breast cancer occur in these countries. Furthermore, while survival rates for breast cancer are around 80% in developed countries, this rate decreases to 60% in middle-income and to 40% in low-income countries due to lack of early detection programs leading to diagnoses in late stages, where 80% of these tumors are incurable [6].

In the middle- and low-income countries, mammography and other expensive and technologically complicated methods are unattainable due to high costs and shortage of trained personnel. Thus, there is a critical need for affordable, portable and accurate means of detecting and diagnosing breast cancer at early stages in the field by basic-level healthcare workers. Development of such technologies has the potential to expedite the solution for the growing health problem to prevent increasing death and disability among women especially in low- and middle-income countries. This critical need has given rise to the study of discovering biomarkers for breast cancer risk which should be easily obtained from patients and quantitatively analyzed by low-cost devices.

Currently, there are a handful of biomarkers that are used in the clinic for breast cancer diagnosis. These biomarkers identify proteins overexpressed in certain subtypes of breast cancers and help clinicians' treatment planning. Up to date, just a few breast cancer biomarkers demonstrated clinical utility, including estrogen receptor alpha (ERα), progesterone receptors (PR), human epidermal growth factor receptor 2 (HER2) in order to predict effectiveness of systemic therapy [7] and the Oncotype DX-21 gene to predict benefit of chemotherapy [8]. Studies evaluating other predictive

biomarkers are in progress for breast cancer type 1 protein (BRCA1) [9], breast cancer type 2 protein (BRCA2) [10], circulating tumor cells (CTCs) [11], HER2-positive [12], topoisomerase II alpha (TOP2A) [13] (in subjects with HER2 over-expression) and HER2 (when is negative in tumors but is positive in the CTCs) [14]. Yet, we still lack reliable, persistent biomarkers of breast cancer risk. Lack of such biomarkers prevents establishment of reliable screening or prevention programs.

In addition to the identified biomarkers, liquid biopsy methods supported with machine learning have been used for the detection of different cancer types [15]. However, a method relying on detection of metabolites and proteins has not been demonstrated systematically in a large clinical study.

To address this critical need, we systematically evaluated metabolites and proteins in plasma to identify potential biomarkers for breast cancer risk that can be utilized with minimally invasive, affordable, portable, and accurate screening devices. In this study, our focus is on liquid biopsy samples from plasma that have the potential to provide simple and minimally invasive information for diagnostic decisions. Unlike the current methodologies that often require costly and intensive analysis of specific metabolites or proteins, sophisticated test kits and user dexterity, we developed a pipeline and showed it is efficient to analyze liquid biopsy samples to detect blood biomarkers and diagnose cancer. This pipeline paves the way towards developing the aforementioned screening devices to be used in the field by basic level healthcare workers in low-resource environments.

Apart from discovering new biomarkers, detecting them in different environments has long been an important topic for scientists because it could pave the way for early diagnosis of several diseases. Specifically, finding and detecting cancer biomarkers are significantly important because of the high mortality rate of the cancer without early diagnosis. Currently, most biomarker detection is processed at dedicated laboratories using big, automated and expensive analyzers, which increases waiting time for the results and costs for the analyses. Portable, faster and cheaper devices are highly desired to replace these big and time-consuming analyzers and also to make biomarker

detection process takes place at patient's bedside. So far, several types of sensors have been used like electrochemical, acoustic, micromechanical, optical etc. to make miniaturized biomarker measurement setups. We proposed a FP interferometer-based spectroscopic biomarker analysis setup to quantitatively detect specific molecules in their environments. Our proposed setup was considered as a cheap and portable alternative to common Fourier transform infrared spectroscopy (FTIR) which is a large and expensive device.

The thesis is structured as follows: In Chapter 2, the background of breast cancer detection techniques, prognostic biomarker discovery studies and utilization of them for detecting cancer risk is discussed. In Chapter 3, properties and preparation of the human blood plasma dataset, on which all the statistical and inference studies are conducted, is introduced to the readers.

In Chapter 4, quantitative analysis of plasma metabolomics and proteomics, which includes statistical analysis and widely-used machine learning techniques, to find robust and discriminative biomarkers is explained. More precisely, the blood serum data is preprocessed to ensure the quality of subsequent analyses, then Student's t-test is used as the main feature selection method to determine the importance of each molecule and the top 20 most important molecules undergo further feature elimination based on pairwise dependencies. Finally, the performance of logistic regression classifier on the remaining molecules is measured and their biological relevance is compared. Several biomarkers of breast cancer risk are identified. Some of these biomarkers are validated using further biological inspection. These biomarkers can be used to develop low-cost screening methods towards early diagnosis and hence decreased mortality due to breast cancer. Our findings in this chapter provide the knowledge basis needed to proceed in this direction.

In Chapter 5, a custom spectroscopic device system, which can be utilized to detect the biomarkers found by the quantitative analysis, is introduced. More precisely, sensor hardware development parts of the project in which we construct a spectroscopy setup to quantitatively detect certain molecules in blood serum samples is covered.

Fabry-Perot interferometer and IR source are placed in a portable 3D printed setup to make measurement process easier. Then, CaF_2 windows, on which we would like to drop samples, are processed in clean room using microfabrication processes like photolithography and material development. In addition, problems with reconstruction of spectroscopic signal obtained by the custom device, limitations and potential solutions are discussed.

In Chapter 6, three pillars of this study, statistical analysis, spectroscopic device building and reconstruction of the spectral signals are discussed together to better show the big picture. Additionally, potential future directions for ill-posed problem of spectral signal reconstruction, improvements for a more robust and reliable measurement setup are argued.

2. BACKGROUND

2.1. Screening Methods for Early Detection of Breast Cancer

The purpose of the screening is to evaluate patients during the early stages of the breast cancer. As stated in Chapter 1, there are various techniques utilized for the early detection of breast cancer to reduce the mortality and morbidity rates from this disease. Each one has its own advantages and limitations. First, we have screening methods like mammography, ultrasound imaging and magnetic resonance imaging (MRI). These screening methods aim to diagnose the breast cancer at early stages and enable more effective treatment.

2.1.1. Mammography

Mammography is the primary screening and diagnostic tool for breast cancer but it is expensive since it should be carried out frequently and requires trained labour. One of the most important disadvantage of mammography is the over-diagnosis [16]. Also, it is not prognostic which means that it can only detect the breast cancer when it is already developed in the body. Although it is smaller relative to the risk of breast cancer, there is a radiation risk which especially needs to be taken into account for younger patients and/or in the case of repetitive exposures.

2.1.2. Ultrasound

Breast screening with ultrasound is used as a supplementary tool to the standard mammography technique. Individuals who have denser breast tissue are recommended for additional screening to reveal a possible tumor masked under the dense tissue [17]. Advantages of ultrasound examination includes lack of ionized radiation, being relatively inexpensive procedure and wide availability. However, ultrasound alone is not sufficient for breast cancer screening because of lower sensitivity to microcalcifications in the breast that may be an indicator of breast cancer [18].

2.1.3. Magnetic Resonance Imaging (MRI)

MRI is a promising screening method used to detect breast cancer because of it's good soft tissue contrast and high sensitivity especially with the use of contrast agents [19]. Similar to the ultrasound, MRI is mainly used as a supplementary tool to the mammography or ultrasound imaging. The procedure mostly performed when the patient is diagnosed with cancer but the size of the diagnosed tumor or potential other lesions in the breast are need to be evaluated [20]. The limitations of MRI are high costs of device and procedure and possibility of false positive inference and unessential biopsies [21].

2.2. Biomarker Based Methods for Early Detection of Breast Cancer

As described in the previous section, each one of the breast cancer screening method has its own limitations. Even mammography, most prevalently used screening method, suffers from false positive or over-diagnosis and insufficient sensitivity [22]. Therefore, it is necessary to invest on predictive biomarkers for breast cancer since they are less complex and it is relatively easier to acquire serum samples.

Statistical analysis methods like Student's t-test is commonly used to identify biomarkers. Although it gives information about the ability of a biomarker to discriminate between healthy and susceptible patients, it alone is not sufficient and machine learning classification methods, mainly logistic regression, are used to identify biomarkers. Area under curve (AUC) value of receiver operating characteristic (ROC) analysis resulting from the classification operations on these biomarkers are commonly used as an indicator for the discriminative capacity of a single molecule or a set of molecules. In [23], logistic regression was performed on predictors consisting of serum levels of several molecules, but they did not report any confidence interval for that AUC value and did not split the data into training and test tests. In [24], they used Student' t-test and its non-parametric equivalence (Mann-Whitney U-test) to find potential biomarkers, but the lower bound of their reported confidence intervals were dramatically low. This showed that those biomarkers are not robust, and they also did not split the

training and test sets. Several studies [22, 25–28] also used these methods to identify potential biomarkers but did not utilize a train-set split for their datasets. Training (model building) and testing on the same dataset is not a good practice in machine learning as the model is likely to overfit to the data. This results in artificially high predictive rates, in other words, low generalization, which refers to poor applicability of the model to unseen data.

As stated in Chapter 1, [15] recently demonstrated the capability of detecting eight different cancer types including breast cancer using circulating tumor DNA (ctDNA) and protein biomarkers. They reported a remarkable sensitivity values greater than 95% for ovary and liver cancers. However, the reported sensitivity for breast cancer is rather low at 33%.

Previously, in 2017, [22], studied pre-diagnostic samples from the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS). Serum samples were taken from 239 women who were diagnosed with invasive ductal carcinoma (IDC) in breast, months to years after sample donation. These patients were post-menopausal women with ages ranging from 50 to 74, who were healthy cases at the moment of recruitment but that later developed breast cancer. Hence this group studied the ability of several (9) serum markers to detect breast cancer cases before these patients were diagnosed. They studied cancer antigen (CA) 15-3, chemokine ligand 5 (CCL5), osteopontin (OPN), plasminogen activator inhibitor-1 (PAI-1), secretory leukocyte protease inhibitor (SLPI), heat shock protein 90A (HSP90A), insulin-like growth factor-binding protein 3 (IGFBP3), apolipoprotein C-I (APOC1) and pappalysin-1 (PAPPA). They concluded that only 3 out of the 9 serum markers, (CA 15-3, PAI-1 and HSP90A) were potential prognostic biomarkers. Those analyses were performed using a limited panel of proteins, however, in our analysis, we characterized more than 300 proteins and metabolites in plasma and used a total of 181 molecules consisting of the most of these metabolites and proteins.

2.3. Measurement Tools for Biomarker Based Methods for Early Detection of Breast Cancer

So far, several types of sensors have been used like electrochemical [29], acoustic [30], micromechanical [31], optical [32] etc. to make miniaturized biomarker measurement setups. Among these methods, optical sensors are promising because of their superior features such as the speed and robustness. There are also other methods such as mass spectrometry investigated for the direct analysis of serum proteins [33, 34].

3. BLOOD PLASMA DATA

In this study, potential prognostic biomarkers for breast cancer risk are extracted from plasma metabolomics and proteomics data using statistical and discriminative power analyses. Therefore, one of the main pillars of this study is the preparation and properties of plasma metabolomics and proteomics data.

3.1. Patients and Plasma Samples

All studies were approved by the Indiana University Institutional Review Board (IRB protocol number 1011003097 and 1607623663). All research was carried out in compliance with the Helsinki Declaration. Donors provided broad consent for the use of their specimens in research. The consent document informed the donor that the donated specimens and medical data would be used for the general purpose of helping to determine how breast cancer develops. It was explained in the consent that the exact laboratory experiments were unknown at the time of donation, and that proposals for use of the specimens would be reviewed and approved by a panel of independent researchers before specimens and/or data were released for research purposes. Hematoxylin and eosin stained sections of the Formalin-Fixed Paraffin-Embedded (FFPE) tissue of the identified donors were reviewed by pathologist to confirm the absence of histological abnormalities.

Two different datasets, one with 39 susceptible and 36 susceptible and the other with 10 susceptible and 11 healthy individuals, were included in this study. In the first dataset, 22 out of 39 susceptible and 23 out of 36 healthy individuals were in postmenopausal status and remaining ones were in premenopausal status. Among susceptible individuals of the first dataset, 9 had body mass index (BMI)$<$25, 10 had 25$<$BMI$<$30 and 20 had BMI$>$30 while for healthy individuals, 8 had BMI$<$25, 16 had 25$<$BMI$<$30 and 12 had BMI$>$30. In the second dataset, 7 out of 10 susceptible and 8 out of 11 healthy individuals were postmenopausal status and remaining ones were premenopausal. Among susceptible individuals of the second dataset, one had

BMI<25, six had 25<BMI<30 and three had BMI>30 while for healthy individuals, two had BMI<25, six had 25<BMI<30 and three had BMI>30.

3.2. Plasma Preparation

Blood was drawn into the plasma separator tube (Vacutainer Venous Blood Collection Tubes: Serum Separation Tube, Plasma Separation Tube, Fisher Scientific cat. 0268396) and gently mixed by inverting the tube 5 times. 45 minutes (±10 minutes) after the blood had been drawn, the plasma separator tube was placed into a mini-centrifuge (Eppendorf Centrifuge 5702) and centrifuged at 1200 relative centrifugal force (RCF) for 10 minutes at room temperature. A repeater pipette was used to aliquot 600 μL of the plasma into each of five cryogenic vials. Samples were stored at -80 °C until use.

3.3. OLINK Protein Biomarker and Whole Metabolite Profiling Assays

All the samples from human studies were handled and analyzed in accordance with UIUC IRB protocol number 06741. 10 μL of plasma samples from Koman Tissue Bank were submitted to OLINK biosciences for cancer and inflammation protein biomarker analysis. 50 μL of plasma samples were submitted to the Metabolomics Center at University of Illinois Urbana-Champaign (UIUC). Gas chromatography (GC) and mass spectrometry (MS) whole metabolite profiling was performed to detect and quantify the metabolites by using GC/MS analysis. Metabolites were extracted from 50 μL of plasma according to Agilent Inc. application notes. The hentriacontanoic acid was added to each sample as the internal standard prior to derivatization. Metabolite profiles were acquired using an Agilent GC/MS system (Agilent 7890 gas chromatograph, an Agilent 5975 mass selective detector, and an Hewlett-Packard 7683B autosampler). The spectra of all chromatogram peaks were evaluated using the Automated Mass Spectral Deconvolution and Identification System (AMDIS) 2.71 and a custom-built database with 460 unique metabolites. All known artificial peaks were identified and removed prior to data analysis. To allow the comparison between samples, all data were normalized to the internal standard in each chromatogram.

3.4. Stem Cell Factor (SCF) Quantification Using Enzyme-Linked Immunosorbent Assay (ELISA)

Serum samples were collected and stored at -80° until the time of assay. We used an ELISA kit for SCF (Sigma-Aldrich, catalog number RAB0330). Samples were diluted 2 fold per suggestion from the manufacturer. The antibody concentrate was diluted 100-fold with 1x Diluent Buffer. To prepare the horseradish peroxidase (HRP)-Streptavidin concentrate, the vial was spin and diluted 400 times with 1x Diluent Buffer. A 50 ng/mL stock solution was used to make the standard curve: 2000 pg/mL, 666.7 pg/mL, 222.2 pg/mL, 74.07 pg/mL, 24.69 pg/mL, 8.23 pg/mL and 2.74 pg/mL.

The human SCF antibody pre-coated ELISA wells were filled with 100 μL of either serially diluted standard protein and serum samples. After 2.5 hours incubation with gentle shaking at room temperature, 100 μL of 1x SCF Biotinylated Detection Antibody were added to the wells. After one hour incubation with shaking at room temperature, the solution was discarded and the wells were washed 4x using 300 μL wash buffer solution. Final wash was aspirated and plates were inverted to remove any remaining was buffer. Then, 100 μL of prepared HRP-Streptavidin solution was added to each well, and incubated for 45 minutes at room temperature with gentle shaking. The solution was discarded and washed 4x as described previously. 100 μL of ELISA colorimetric TMB reagent was added to each well and incubated for 30 minutes at room temperature in the dark. After this, 50 μL of stop solution was added to each well. Immediately after color development, the OD values were measured at 450 nm using Cytation 5 Cell Imaging Multi-Mode Reader (Biotek Inc.) and SCF concentrations were calculated from specific calibration curves prepared with known standard solutions. Diluent buffer served as blank and the OD of these wells was subtracted from the values.

4. STATISTICAL AND BIOLOGICAL ANALYSIS

4.1. Preprocessing

Initially, plasma levels of 295 different molecules for the first dataset and 339 different molecules for the second dataset were measured for the individuals but some molecules had missing values for some individuals. Also, some molecules were not measured for both datasets and all these molecules were excluded from the analysis Therefore, we analyzed a total of 181 different molecules, consisting of metabolites and proteins, which have plasma level values for every subject in both datasets.

We normalized all individuals' plasma data in each dataset with respect to the healthy individuals' data in the respective dataset to factor out potential differences in data acquisition. More specifically, we performed the following procedure separately for both datasets. For each individual k and molecule i, we subtracted the mean measurement of that molecule in healthy individuals $\mu_{i_{healthy}}$ from all individuals' measurements $x_i^{(k)}$ and divided this difference by the standard deviation of that molecule's measurements in healthy individuals $\sigma_{i_{healthy}}$ to find a sort of normalized molecule measurements $\hat{x}_i^{(k)}$ as follows:

$$\hat{x}_i^{(k)} = \frac{x_i^{(k)} - \mu_{i_{healthy}}}{\sigma_{i_{healthy}}} \tag{4.1}$$

$$\mu_{i_{healthy}} = \frac{\sum_{k \in \{healthy\}} x_i^{(k)}}{N_{healthy}} \tag{4.2}$$

$$\sigma_{i_{healthy}} = \sqrt{\frac{\sum_{k\in\{healthy\}} \left(x_i^{(k)} - \mu_{i_{healthy}}\right)^2}{N_{healthy}}} \tag{4.3}$$

where $N_{healthy}$ is the number of healthy individuals in respective dataset. Thus by applying the transformations above to both datasets, we converted each single measurement to a z-score which describes the deviation of that measurement from the mean of healthy individual's, in terms of the standard deviation among healthy individuals. As the final step, we merged two datasets, which were normalized with respect to their own healthy individuals, and obtained a combined dataset with 49 susceptible and 47 healthy individuals.

4.2. Molecule Ranking and Elimination

4.2.1. Molecule Ranking by Student's t-Test

To assess the molecules' importance, we used Student's t-test to compare the two classes (healthy and susceptible) for each molecule with null hypothesis being that the measurements in the two classes come from the same distribution. We ranked the molecules based on their p-values calculated by t-test and short-listed the top ranking 20 molecules with the lowest p-values for further elimination. These top ranking molecules can be seen in Table 4.1.

The selected molecules by their p-values underwent three different further elimination processes separately. These three different methods includes elimination by logistic regression (LR), correlation analysis and paired t-test analysis. All of them have different approaches for giving importance to each molecule so we would like to evaluate the molecules by different measures.

Table 4.1: First 20 molecules ranked by their p-values.

Rank	**Student's T-Test**	**p-value**
1	SCF	0.0091
2	MAD HOMOLOG 5	0.0215
3	FGF-5	0.0288
4	FASLG	0.0298
5	MMP-10	0.0328
6	PPY	0.0388
7	XPNPEP2	0.0473
8	FGF-21	0.0508
9	CXL17	0.0526
10	MCP-3	0.064
11	ESM-1	0.067
12	HK11	0.0791
13	TRAIL	0.0801
14	FGF-BP1	0.1245
15	EN-RAGE	0.1252
16	C15:0	0.1356
17	TNFB	0.1746
18	CTSV	0.1797
19	ADA	0.1841
20	CD160	0.1887

4.2.2. Molecule Elimination by Logistic Regression

First, we manually eliminated molecules based on their LR classification performance assessed using AUC of ROC curves. Logistic regression classifier basically constructs affine transformation below followed by a sigmoid function on the input variables (features) $\hat{x}_i^{(k)}$ with weights w_i and a bias term β.

$$\hat{y}^{(k)} = \sigma(w_1\hat{x}_1^{(k)} + \ldots + w_i\hat{x}_i^{(k)} + \ldots + w_D\hat{x}_D^{(k)} + \beta) \tag{4.4}$$

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{4.5}$$

where D is the number of input features (dimensionality of features), the variables $\hat{x}_i^{(k)}$ corresponds to the normalized measurement of molecule i for each sample k and $\hat{y}^{(k)} \in [0,1]$ is the predicted probability assignment to the label of corresponding sample k (Healthy = 0, Susceptible = 1) with true label $y^{(k)} \in \{0,1\}$. Using the result of the transformations above and the true labels, logistic regression classifier defines the following likelihood function to maximize:

$$L(\mathbf{y}, \hat{\mathbf{y}}) = \prod_{k=1}^{N} P_{\boldsymbol{\theta}}\big(y^{(k)}|\hat{\mathbf{x}}^{(k)}\big) = \prod_{k=1}^{N} \Big(\hat{y}^{(k)}\Big)^{y^{(k)}} \Big(1-\hat{y}^{(k)}\Big)^{1-y^{(k)}} \tag{4.6}$$

where

$$\mathbf{y} = [y^{(1)}, \ldots, y^{(k)}, \ldots, y^{(N)}] \tag{4.7}$$

$$\hat{\mathbf{y}} = [\hat{y}^{(1)}, \ldots, \hat{y}^{(k)}, \ldots, \hat{y}^{(N)}] \tag{4.8}$$

$$\hat{\mathbf{x}}^{(k)} = [\hat{x}_1^{(k)}, \ldots, \hat{x}_i^{(k)}, \ldots, \hat{x}_D^{(k)}] \tag{4.9}$$

In addition, N is the total number of individuals (samples) and the parameters $\boldsymbol{\theta}$ include the weights $\mathbf{w} = [w_1, \ldots, w_i, \ldots, w_D]$ and bias β. Logistic regression classifier basically maximizes the likelihood in Equation 4.6 by finding the optimal parameters $\boldsymbol{\theta}$ using the measurements of samples.

As stated, performance of different molecules were assessed using AUC of ROC. Using the optimum parameters $\boldsymbol{\theta}$, for each sample k, LR classifier produces $P_{\boldsymbol{\theta}}\big(C_1|\hat{\mathbf{x}}^{(k)}\big)$ for positive class (Susceptible = 1) and $P_{\boldsymbol{\theta}}\big(C_0|\hat{\mathbf{x}}^{(k)}\big) = 1 - P_{\boldsymbol{\theta}}\big(C_1|\hat{\mathbf{x}}^{(k)}\big)$ for negative class (Healthy = 0). Assume we choose a sample k "positive" if the following equation holds: $P_{\boldsymbol{\theta}}\big(C_1|\hat{\mathbf{x}}^{(k)}\big) > \phi$. Then by changing this threshold $\phi \in [0, 1]$, we can get different prediction accuracies. At this point, we should define true positives (tp), false positives (fp), true negatives (tn) and false negatives (fn) as below.

Table 4.2: Demonstration of tp, fp, tn and fn.

	Predicted Class		
True Class	Positive (Susceptible)	Negative (Healthy)	Total
Positive (Susceptible)	tp	fn	tp + fn
Negative (Healthy)	fp	tn	fp + tn
Total	tp + fp	fn + tn	N

As we increase the threshold ϕ, we can decrease fp but in the meantime tp also decreases [35]. The same relation holds for tn and fn rates while we decrease ϕ. Therefore, these numbers change by changing the threshold ϕ. ROC curve basically shows the relation between tp-rate $\left(\frac{\text{tp}}{\text{tp+fn}}\right)$ and fp-rate $\left(\frac{\text{fp}}{\text{fp+tn}}\right)$ as threshold changes. As previously stated, we can measure the performance of a classifier by calculating AUC of ROC. This is because we always want our classifier to predict the true classes as accurate as possible and therefore it should give high tp-rate and low fp-rate. This means that better classifiers have ROC curves closer to the upper left corner and so the AUC of ROC is higher for better classifiers. Thus, we can easily use AUC of ROC as an indicator of classifier performance. An example can be seen in Figure 4.1.

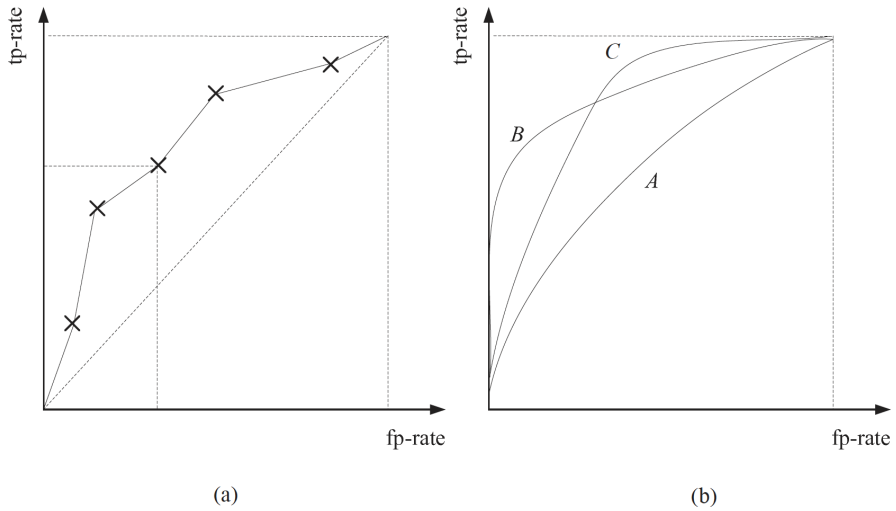


Figure 4.1: (a) Typical ROC curve. At each point on this ROC curve, we have different thresholds ϕ and therefore get different tp- and fp-rates. (b) Different classifiers produce different ROC curves. We prefer the highest AUC. [35]

We checked the AUC of ROC performance change of LR classifier, using the best-ranking 20 molecules found in the previous Section 4.2.1, while increasing the number of molecules used, starting from the top ranking one, as features and removed the ones that decreased the performance. The remaining molecules after the elimination can be seen under "LR Classification" in Table 4.5.

4.2.3. Molecule Elimination by Correlation and Paired t-Test Analysis

In this part, we assessed the inter-dependency of the top ranking 20 molecules found in Section 4.2.1 by using correlation and paired t-test analysis. We initialized a "selected molecules" list (S-list) with the top-ranking molecule and the "unselected molecules" list (U-list) with the remaining 19 molecules. All molecules in the U-list are passed through the paired t-test with each one of the molecules in the S-list, with null hypothesis being that both come from the same distribution. The molecules in the U-list with p-value higher than 0.05 are eliminated, while the one with the lowest p-value (< 0.05) is moved to the S-list. The procedure is repeated until U-list is empty. Finally, we applied correlation analysis instead of paired t-test to eliminate the molecules in the above described iterative approach. In each iteration, we eliminated the molecules

in U-list with a Pearson's correlation coefficient (pCC) higher than 0.5 with anyone of the molecules in S-list.

Table 4.3 shows the p-values of the paired t-test analysis for each pair of molecules among the top 20 molecules ranked by applying the Student's t-test to all 204 molecules in our dataset. The Student's t-test was used to rank the molecules with respect to their (healthy vs susceptible) discriminative power, while the paired t-test assesses the pairwise dependencies of the most discriminative 20 molecules. Table 4.4 shows the results of pairwise correlation analysis for the same top-ranking 20 molecules. The molecule pairs with high p-value (p$>$ 0.05) and/or correlation coefficient (pCC$>$ 0.5) indicate a redundancy within the pair, hence of the one of the molecules in that pair can be discarded, which we utilized for molecule elimination. The remaining molecules after the elimination by both correlation and paired t-test analysis can be seen under "Correlation Analysis" and "Paired T-Test" in Table 4.5.

4.3. Classification on Selected Molecules

We applied LR to different molecule sets selected by different molecule elimination methods in order to classify individuals into classes of healthy and susceptible and assessed molecules' discriminative power. AUC of ROC curves were used as a measure of the discriminative power of the molecules as in Section 4.2.2. Normalized AUC ($\in [0, 1]$) of ROC curves provides a summary statistic of the discriminative power of the selected molecules. As previously explained, the higher the AUC the higher the discriminative power is.

Table 4.5 summarizes the molecule ranking and elimination results. The "Student's t-test" column lists the top ranking most discriminative 20 molecules according to their p-values among 204 molecules. The "LR Classification", "Paired t-test" and "Correlation Analysis" columns list the molecules selected from these 20 top molecules by applying the iterative molecule elimination procedures using LR performance based on AUC values, paired t-test and correlation analysis, respectively, as described in Section 4.2.1, 4.2.2 and 4.2.3.

Table 4.3: Paired t-test p-values between each pair of 20 most important molecules selected by their p-values.

	SCF	MAD HOMOLOG 5	FGF-5	FASLG	MMP-10	PPY	XPNPEP2	FGF-21	CXL17	MCP-3	ESM-1	HK11	TRAIL	FGF-BP1	EN-RAGE	C15:0	TNFB	CTSV	ADA	CD160
SCF	0	0.6963	0.0004	0.0006	0.3286	0.0017	0.361	0.003	0.2659	0.189	0.0003	0.1552	0.1309	0.1303	0.1191	0.1324	0.0029	0.0045	0.001	0.0015
MAD HOMOLOG 5	0.6963	0	0.0013	0.0014	0.6643	0.0038	0.5633	0.0026	0.4489	0.2985	0.0006	0.3257	0.5298	0.2913	0.354	0.2454	0.008	0.0062	0.0083	0.007
FGF-5	0.0004	0.0013	0	0.6618	0.002	0.7183	0.0018	0.6703	0.0017	0.0027	0.6315	0.0064	0.0033	0.01	0.0027	0.0091	0.4087	0.4174	0.6858	0.4709
FASLG	0.0006	0.0014	0.6618	0	0.0018	0.9554	0.0015	0.9738	0.0034	0.002	0.9137	0.0015	0.0039	0.0112	0.005	0.0067	0.6151	0.548	0.9595	0.6497
MMP-10	0.3286	0.6643	0.002	0.0018	0	0.0033	0.9115	0.0006	0.8086	0.542	0.0073	0.5825	0.8731	0.5308	0.6119	0.417	0.0071	0.0154	0.0169	0.0086
PPY	0.0017	0.0038	0.7183	0.9554	0.0033	0	0.0054	0.9356	0.0018	0.0055	0.8852	0.0063	0.0096	0.0166	0.0124	0.0109	0.6315	0.6427	0.93	0.6801
XPNPEP2	0.361	0.5633	0.0018	0.0015	0.9115	0.0054	0	0.0034	0.9172	0.66	0.0125	0.6878	0.9708	0.6209	0.6971	0.4563	0.0108	0.0055	0.0188	0.0214
FGF-21	0.003	0.0026	0.6703	0.9738	0.0006	0.9356	0.0034	0	0.0014	0.0029	0.9489	0.006	0.0063	0.013	0.0134	0.0096	0.7003	0.6644	0.9805	0.7499
CXL17	0.2659	0.4489	0.0017	0.0034	0.8086	0.0018	0.9172	0.0014	0	0.7384	0.0038	0.7438	0.9536	0.6714	0.7648	0.5413	0.0172	0.026	0.019	0.0126
MCP-3	0.189	0.2985	0.0027	0.002	0.542	0.0055	0.66	0.0029	0.7384	0	0.0089	0.9712	0.7017	0.9032	0.9911	0.788	0.0183	0.017	0.0211	0.0241
ESM-1	0.0003	0.0006	0.6315	0.9137	0.0073	0.8852	0.0125	0.9489	0.0038	0.0089	0	0.0032	0.0061	0.013	0.0112	0.0289	0.7453	0.7257	0.9715	0.7768
HK11	0.1552	0.3257	0.0064	0.0015	0.5825	0.0063	0.6878	0.006	0.7438	0.9712	0.0032	0	0.7392	0.8734	0.9815	0.7421	0.0202	0.0234	0.037	0.0087
TRAIL	0.1309	0.5298	0.0033	0.0039	0.8731	0.0096	0.9708	0.0063	0.9536	0.7017	0.0061	0.7392	0	0.6469	0.6816	0.5747	0.0092	0.0274	0.0071	0.0078
FGF-BP1	0.1303	0.2913	0.01	0.0112	0.5308	0.0166	0.6209	0.013	0.6714	0.9032	0.013	0.8734	0.6469	0	0.9125	0.8767	0.0416	0.0442	0.0429	0.0441
EN-RAGE	0.1191	0.354	0.0027	0.005	0.6119	0.0124	0.6971	0.0134	0.7648	0.9911	0.0112	0.9815	0.6816	0.9125	0	0.8043	0.0374	0.025	0.0115	0.0327
C15:0	0.1324	0.2454	0.0091	0.0067	0.417	0.0109	0.4563	0.0096	0.5413	0.788	0.0289	0.7421	0.5747	0.8767	0.8043	0	0.0409	0.0464	0.0725	0.0502
TNFB	0.0029	0.008	0.4087	0.6151	0.0071	0.6315	0.0108	0.7003	0.0172	0.0183	0.7453	0.0202	0.0092	0.0416	0.0374	0.0409	0	0.969	0.7488	0.9447
CTSV	0.0045	0.0062	0.4174	0.548	0.0154	0.6427	0.0055	0.6644	0.026	0.017	0.7257	0.0234	0.0274	0.0442	0.025	0.0464	0.969	0	0.7132	0.9266
ADA	0.001	0.0083	0.6858	0.9595	0.0169	0.93	0.0188	0.9805	0.019	0.0211	0.9715	0.037	0.0071	0.0429	0.0115	0.0725	0.7488	0.7132	0	0.8146
CD160	0.0015	0.007	0.4709	0.6497	0.0086	0.6801	0.0214	0.7499	0.0126	0.0241	0.7768	0.0087	0.0078	0.0441	0.0327	0.0502	0.9447	0.9266	0.8146	0

Table 4.4: Pearson's correlation coefficient between each pair of twenty most important molecules ranked by their p-values.

	SCF	MAD HOMOLOG 5	FGF-5	FASLG	MMP-10	PPY	XPNPEP2	FGF-21	CXL17	MCP-3	ESM-1	HK11	TRAIL	FGF-BP1	EN-RAGE	C15:0	TNFB	CTSV	ADA	CD160
SCF	1.000	0.136	0.129	0.076	0.318	-0.088	0.030	-0.189	0.215	0.030	0.265	0.261	0.660	0.228	0.415	0.000	0.073	-0.024	0.312	0.223
MAD HOMOLOG 5	0.136	1.000	0.057	0.064	-0.048	-0.119	0.148	0.021	0.323	0.217	0.305	0.208	0.244	0.118	0.132	0.017	0.012	0.083	0.089	0.082
FGF-5	0.129	0.057	1.000	0.210	0.047	0.150	0.123	0.104	0.151	0.116	0.067	-0.048	0.113	-0.073	0.233	-0.027	0.184	0.049	0.219	0.161
FASLG	0.076	0.064	0.210	1.000	0.079	0.169	0.179	0.192	0.037	0.165	0.149	0.246	0.127	-0.068	0.154	0.075	0.285	0.384	0.120	0.446
MMP-10	0.318	-0.048	0.047	0.079	1.000	-0.002	0.156	0.320	0.323	0.281	-0.087	0.264	0.294	0.076	0.143	0.097	0.159	-0.042	0.043	0.158
PPY	-0.088	-0.119	0.150	0.169	-0.002	1.000	-0.047	0.085	0.199	0.012	-0.006	0.022	-0.065	-0.147	-0.034	0.003	0.089	-0.175	0.035	0.173
XPNPEP2	0.030	0.148	0.123	0.179	0.156	-0.047	1.000	0.110	0.168	0.150	-0.169	0.188	-0.030	0.004	0.118	0.227	0.138	0.281	0.083	-0.018
FGF-21	-0.189	0.021	0.104	0.192	0.320	0.085	0.110	1.000	0.276	0.200	-0.124	0.092	0.101	0.003	0.005	0.103	0.016	0.038	0.065	0.097
CXL17	0.215	0.323	0.151	0.037	0.323	0.199	0.168	0.276	1.000	0.129	0.164	0.320	0.198	0.092	0.141	0.131	0.042	-0.096	0.119	0.170
MCP-3	0.030	0.217	0.116	0.165	0.281	0.012	0.150	0.200	0.129	1.000	0.020	0.068	0.214	0.112	0.319	-0.112	0.110	0.133	0.202	0.079
ESM-1	0.265	0.305	0.067	0.149	-0.087	-0.006	-0.169	-0.124	0.164	0.020	1.000	0.269	0.160	0.072	0.117	-0.187	0.075	-0.026	0.233	0.299
HK11	0.261	0.208	-0.048	0.246	0.264	0.022	0.188	0.092	0.320	0.068	0.269	1.000	0.176	0.187	0.047	0.130	0.112	0.071	0.007	0.352
TRAIL	0.660	0.244	0.113	0.127	0.294	-0.065	-0.030	0.101	0.198	0.214	0.160	0.176	1.000	0.134	0.435	-0.066	0.280	-0.009	0.379	0.333
FGF-BP1	0.228	0.118	-0.073	-0.068	0.076	-0.147	0.004	0.003	0.092	0.112	0.072	0.187	0.134	1.000	-0.055	0.008	-0.017	-0.034	0.062	0.006
EN-RAGE	0.415	0.132	0.233	0.154	0.143	-0.034	0.118	0.005	0.141	0.319	0.117	0.047	0.435	-0.055	1.000	-0.136	0.019	0.165	0.410	0.107
C15:0	0.000	0.017	-0.027	0.075	0.097	0.003	0.227	0.103	0.131	-0.112	-0.187	0.130	-0.066	0.008	-0.136	1.000	0.035	-0.012	-0.162	-0.001
TNFB	0.073	0.012	0.184	0.285	0.159	0.089	0.138	0.016	0.042	0.110	0.075	0.112	0.280	-0.017	0.019	0.035	1.000	0.048	0.204	0.275
CTSV	-0.024	0.083	0.049	0.384	-0.042	-0.175	0.281	0.038	-0.096	0.133	-0.026	0.071	-0.009	-0.034	0.165	-0.012	0.048	1.000	0.250	-0.087
ADA	0.312	0.089	0.219	0.120	0.043	0.035	0.083	0.065	0.119	0.202	0.233	0.007	0.379	0.062	0.410	-0.162	0.204	0.250	1.000	0.054
CD160	0.223	0.082	0.161	0.446	0.158	0.173	-0.018	0.097	0.170	0.079	0.299	0.352	0.333	0.006	0.107	-0.001	0.275	-0.087	0.054	1.000

Table 4.5: Top ranking 20 molecules according to their p-values and the remaining molecules after further elimination by each method: LR classification performance, correlation and paired t-test analysis.

Rank	**Student's T-Test**	**LR Classification**	**Correlation Analysis**	**Paired T-Test**
1	SCF	SCF	SCF	SCF
2	MAD HOMOLOG 5	MAD HOMOLOG 5	MAD HOMOLOG 5	FGF-5
3	FGF-5	FGF-5	FGF-5	
4	FASLG	FASLG	FASLG	
5	MMP-10	MMP-10	PPY	
6	PPY	XPNPEP2	XPNPEP2	
7	XPNPEP2	FGF-21	FGF-21	
8	FGF-21	CXL17	MCP-3	
9	CXL17	MCP-3	FGF-BP1	
10	MCP-3	ESM-1	C15:0	
11	ESM-1	TNFB		
12	HK11	CTSV		
13	TRAIL	CD160		
14	FGF-BP1			
15	EN-RAGE			
16	C15:0			
17	TNFB			
18	CTSV			
19	ADA			
20	CD160			

Using the selected molecules listed in Table 4.5, we performed LR classification using the top-ranking N molecules in each list (column) of Table 3, where N runs from one to the length of the corresponding list. We performed 4-fold cross-validation with 500 iterations. Figure 4.2 (A-D) shows the AUC values with 95% confidence interval (CI) versus N. ROC curves of every molecule group for each list are also shown in Figure 4.2 (A-D). As can be seen in Figure 4.2 (A-D), we achieved AUC value as high as 0.85 (CI: $\pm$0.04) with 13 molecules and also achieved a satisfactory AUC value 0.69 (CI: $\pm$0.03) with as low as 2 molecules.

Figure 4.2 (A) demonstrates LR classification performances (AUC values) using the top-ranking N (1-20) molecules, ranked by their p-values in Table 4.5, and the ROC curves of every AUC value where the bold black line indicates ROC curve of the best performing (the highest AUC value) molecule set. Figure 4.2 (B) shows LR classification performances (AUC values) using the top-ranking N (1-13) molecules selected manually by considering the LR classification performance given in Figure 4.2 (A) and the ROC curves of every AUC value where the bold black line indicates ROC curve of the best performing (the highest AUC value) molecule set. Figure 4.2 (C) shows LR classification performances (AUC values) using the top-ranking N (1-10) molecules selected from the list of 20 molecules, ranked by Student's t-test in Table 4.5, by iterative elimination using pairwise Pearson correlation coefficients of features in Table 4.4 ($|pCC|=0.5$ is the significance threshold) and the ROC curves of every AUC value where the bold black line indicates ROC curve of the best performing (the highest AUC value) molecule set. (D) gives LR classification performances (AUC values) using the top-ranking N (1-2) molecules selected from the list of 20 molecules, ranked by Student's t-test in Table 4.5, by iterative elimination using paired t-test p-values of features in Table 4.3 ($p=0.05$ is the significance threshold) and the ROC curves of every AUC value where the bold black line indicates ROC curve of the best performing (the highest AUC value) molecule set. Figure 4.2

4.4. Biological Relevance of Selected Molecules

Next, we validated one of the biomarkers, SCF, which was identified with all three feature selection methods. SCF plasma levels were lower in susceptible individuals compared to plasma levels in healthy individuals as in Figure 4.3. We also validated our finding from OLINK analysis using another independent method, ELISA and showed that level of these protein is lower in susceptible individuals as demonstrated in Figure 4.4. Level of the identified biomarker in human plasma samples was compared using unpaired t-test and p-values are reported on both Figure 4.3 and 4.4.

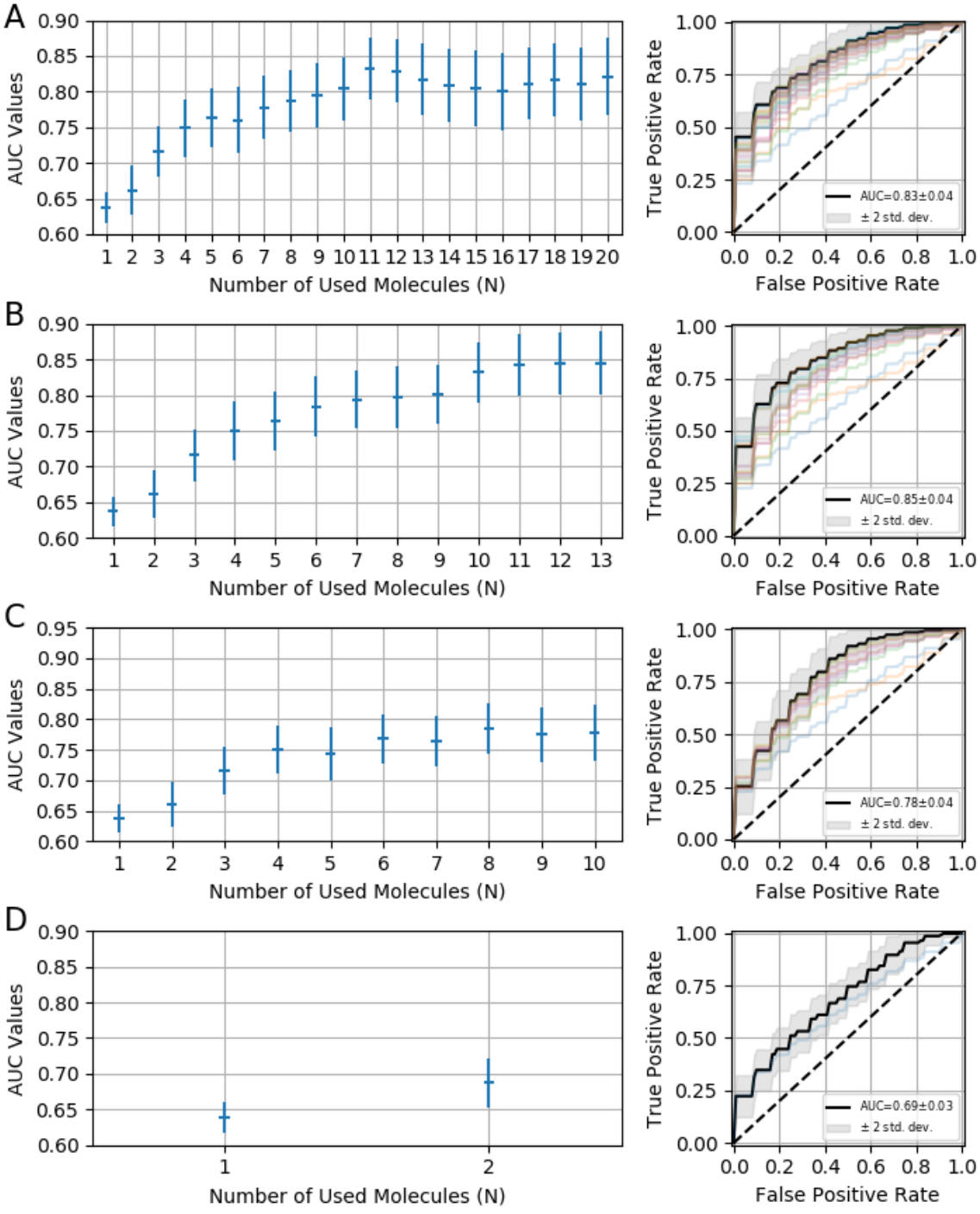


Figure 4.2: (A) Results on top-ranking 20 molecules. (B) Results on the molecules selected by LR classification performance. (C) Results on the molecules selected by correlation analysis. (D) Results on the molecules selected by paired t-test analysis.

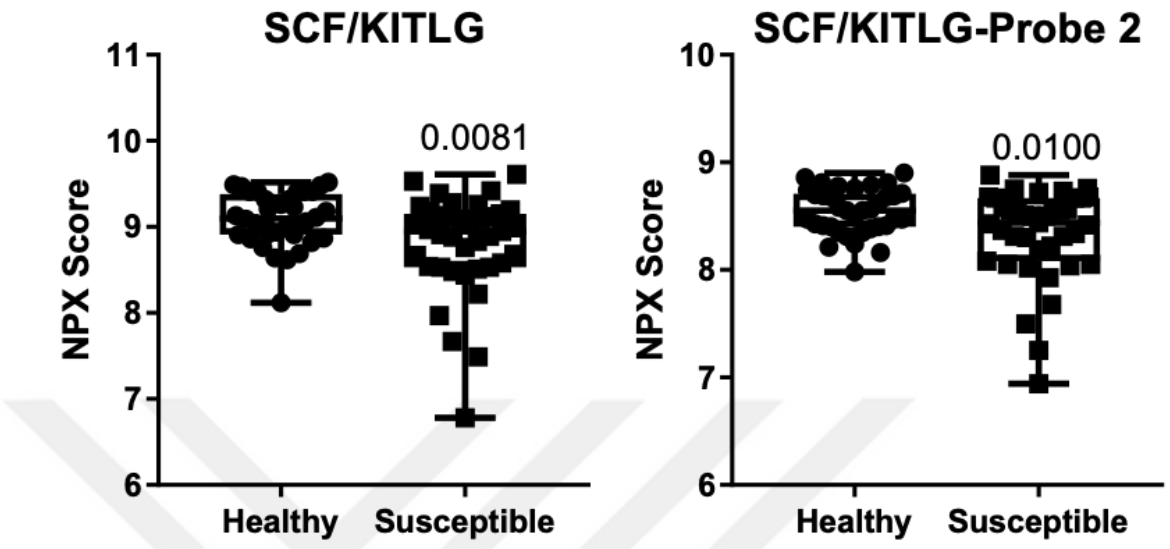


Figure 4.3: Changes in the levels of SCF in healthy and susceptible individuals in the original data obtained by OLINK analysis. There were two probes for SCF in the original analysis.

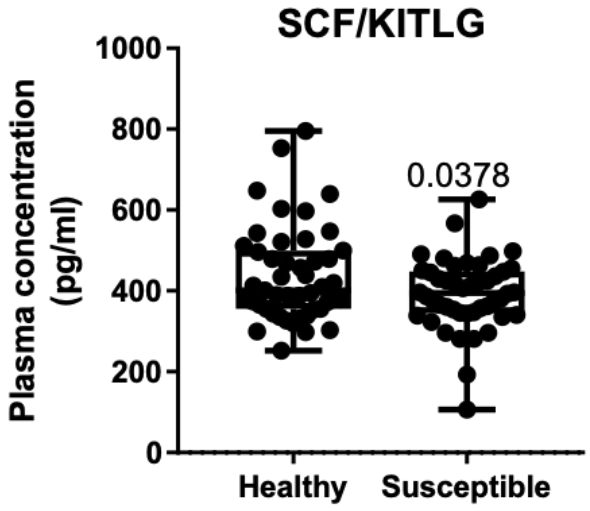


Figure 4.4: Changes in the levels of SCF in healthy and susceptible individuals in additional samples. Results from Figure 4.3 are independently validated using ELISA assay.

4.5. Discussion of Statistical and Biological Analysis

In this chapter, we developed a pipeline to identify plasma biomarkers of breast cancer risk using statistical and machine learning approaches and independently validated one of the identified biomarkers, SCF. A biomarker is characteristic that can be measured to describe that an abnormal or a normal process is taking place within the organism. It can describe different biomolecules such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA), proteins, hormones and chemical modifications [36]. A cancer biomarker can arise due to changes in the DNA (mutations), rearrangements, deletions (missing copies), or the opposite, having extra copies of genes. Whether they are molecules, proteins or genes, they affect various hallmarks of cancer including cell cycle, interfering with cell proliferation, and apoptosis (cell death). In cancer, a biomarker can be linked to a specific molecular pathway. Thus, it can measure the risk of developing cancer, its progression and response to therapy [36].

One potential marker of interest, which we identified is SCF protein. SCF protein is expressed in 53% of breast cancer cell lines [37]. SCF was shown to have a proliferative role in BCK4 cells, and when it is reduced, it decreased estrogen-induced proliferation [38].

5. SPECTROSCOPIC DEVICE SETUP AND SIGNAL RECONSTRUCTION

As stated in Chapter 1, we proposed a new Fabry-Perot (FP) spectroscopy based measurement setup to simply quantify the biomarkers in blood serum sample.

5.1. Measurement Setup

The essential part of the whole biomarker detection setup is our tunable Fabry-Perot interferometers in Figure 5.2 which are commercial and sold by InfraTec GmbH. Because of their tuning capability, these interferometers can change their detection wavelengths in the infrared (IR) region. If we assume that light comes to the interferometer with right angle, the wavelength of a FP interferometer is determined by the following expression:

$$\lambda_m = \frac{2nd}{m} \tag{5.1}$$

where m is the order of interference, n is the refractive index of the medium between the reflectors and d is the distance between reflectors. The tuning system of FP interferometers is illustrated in Figure 5.1. By changing the voltage V_c across two reflectors, the distance between them and so the wavelength that FP interferometer can detect changes piezoelectrically.

Two different interferometers with 3.1-4.4 μm and 5.5-8 μm wavelength bands were used for our setup. The reason behind this choice is that we try to measure in which wavelengths our samples will absorb the light. Absorption wavelengths and amounts show the fingerprints of molecules. Because the molecules measured in our

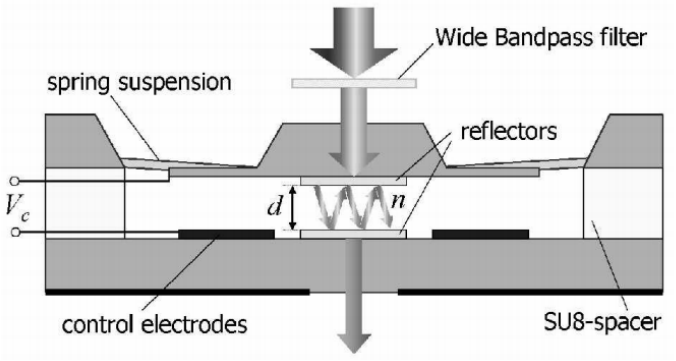


Figure 5.1: Tuning mechanism of FP interferometers sold by InfraTec GmbH [39].

dataset have absorption wavelengths dominantly in this wavelength windows, we chose the FP interferometers with these wavelength windows..

By tuning these interferometers in their respective bands with small steps, we can measure the magnitude of the light passing through the interferometer at each step. Using these measurements at each step, the whole spectroscopy of any measured substance can be obtained.

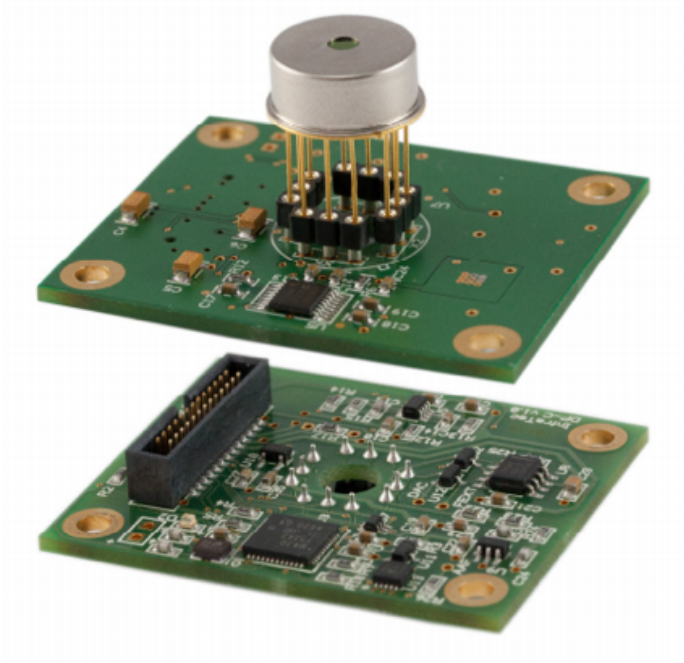


Figure 5.2: Fabry-Perot interferometer and its closed-circuit control board which are both commercially available and sold by InfraTec GmbH.

Because Fabry-Perot interferometers are just sensors which measure the magnitude of the light passing through them, we needed and used the evaluation board in Figure 5.3, sold by InfraTec GmbH. It tunes the interferometers by changing the voltage across the piezoelectric structure inside them, converts analog measurements to digital data by analog to digital converter (ADC) and in addition, control the IR source power, which we used for illuminating samples.

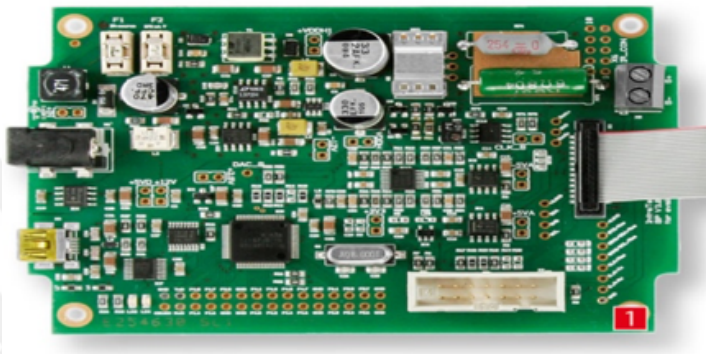


Figure 5.3: Evaluation board for Fabry-Perot interferometers and controller for IR source power.

To take measurements, we need to drop our samples onto a surface and this surface need to be IR transparent in order not to block the light passing through the samples which reaches to the interferometer. We used CaF_2 windows for this purpose because of their high transparency to IR light up to 8 μm and high damage threshold. They were bought from Thorlabs Inc. and its transmission spectra can be seen in 5.4.

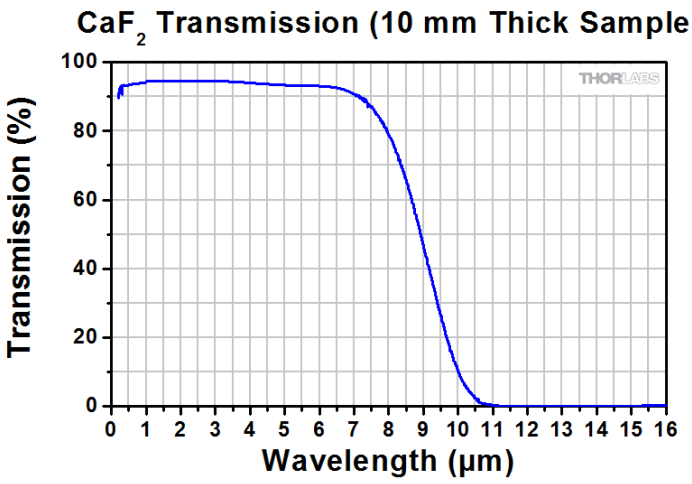
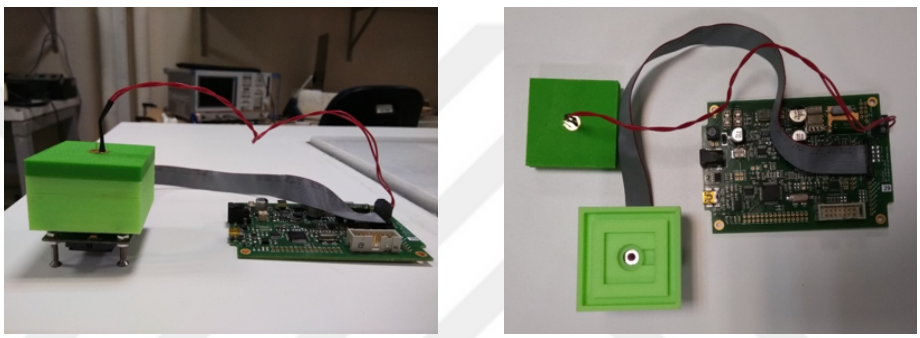


Figure 5.4: Transmission spectra of CaF_2 windows bought from Thorlabs [40].

Overall, the measurement setup in Figure 5.5 was built as follows. IR source, sample on CaF_2 and the FP interferometers were placed linearly in a fabricated setup. IR source and the interferometers were connected to the evaluation board for tuning their parameters. In this setup, IR source illuminated a bare CaF_2 window or a sample on it with broadband light. After this, light had passed through the window, or sample and window, it arrived at the interferometers and its magnitude around the wavelength, to which we tuned our FP interferometers, measured. Then the measured signal reached the evaluation board and it was converted to a usable form.



(a) Side view. (b) Top view.

Figure 5.5: Overall measurement setup which includes FP interferometer, IR source, CaF_2 window and evaluation board connected to the interferometer and IR source.

5.2. Material Development on IR-Transparent Window

Measurements taken by our setup is highly dependent on the sample thickness through which IR light passes before reaching the sensor. To make measurements more robust, we constructed a constant thickness hole on a CaF_2 window by developing and patterning a thin SU-8 layer on the window using photolithography processes in clean room. We used SU-8 2050 which is commercially available and sold by MicroChem Corp. First, we covered our substrate, CaF_2 window, with SU-8 properly. Then we spun the covered window at spin speed of 4000 rpm. We soft baked it at 65 ^oC for 3 minutes. After that, we exposed it to ultraviolet (UV) radiation and baked it again after exposure at 95 ^oC for 6 minutes. Then we developed SU-8 for 5 minutes by using the development solution sold by the same producer MicroChem Corp. Finally, we rinsed and dried it to make it ready for measurements.

5.3. Low Resolution Measurements by Spectroscopic Device

Fabry-Perot interferometers we used have an inherent blurring effect because there is a smoothing kernel at each wavelength which spreads out to the neighbors of that wavelength and adds the effect of these neighbors to the actual result at that magnitude. We measured a chemical sample using our interferometry setup with processed CaF2 windows and also using an FTIR setup with which we obtained ground truths of the samples. Results can be seen in Figure 5.6. As can be seen from the measurement, FP spectra were oversmoothed versions of FTIR spectra. They needed to be reconstructed so that they could give results similar to our ground truth FTIR's.

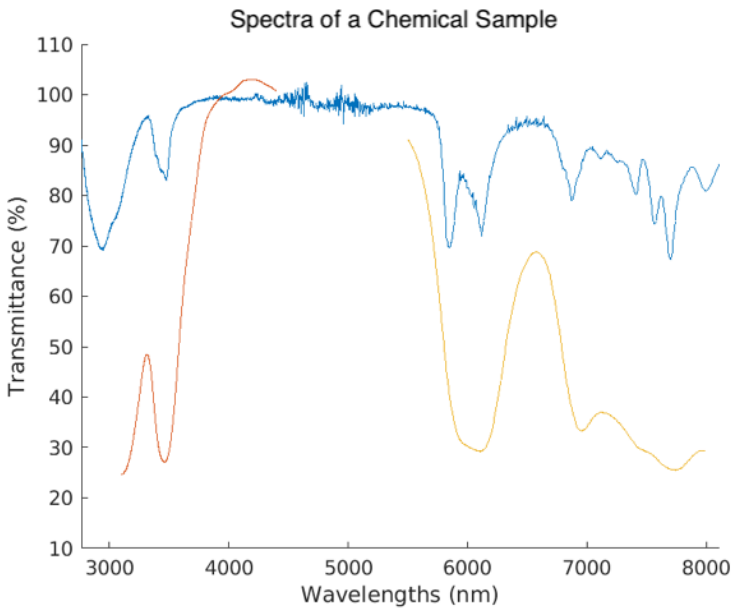


Figure 5.6: Measured FTIR spectrum (ground truth) and FP spectrum from a random chemical sample.

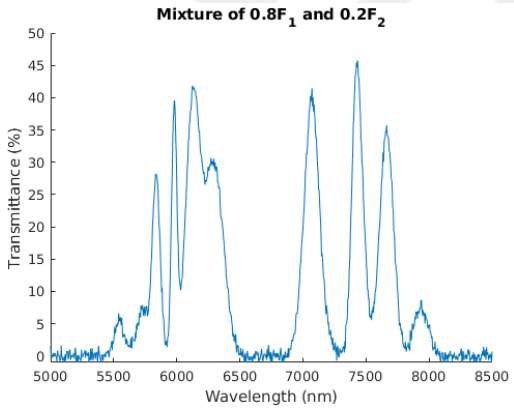
5.4. Spectral Reconstruction by Least Absolute Shrinkage and Selection Operator (LASSO): Simulation Experiments

5.4.1. Spectral Computer Simulations

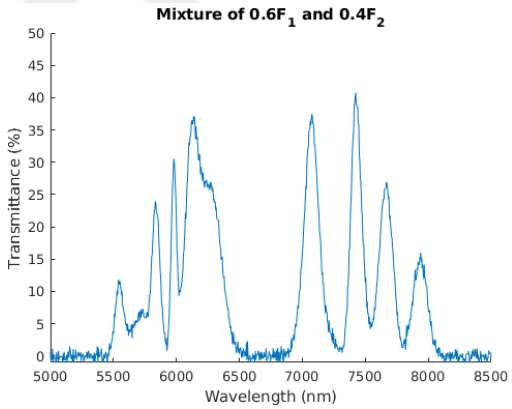
Fabry-Perot interferometers we used have an inherent blurring effect because there is a smoothing kernel at each wavelength which spreads out to the neighbors

of that wavelength and adds the effect of these neighbors to the actual result at that magnitude. To reconstruct actual spectra using the blurred ones, we first generated some ground truth spectra (called as FTIR).

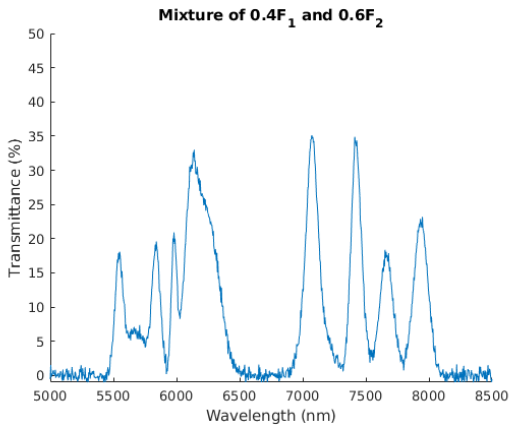
In order to create computer generated FTIR spectra, we created unnormalized Gaussian distributions using random parameters (randomly selected mean, variance and scale in appropriate ranges) centered at 10 random wavelength points between 5.5-8 μm. Then by summing up these 10 distributions, we created a computer-generated spectrum with approximately 10 peaks. By repeating these steps, we created 2 different spectra ($\mathbf{F}_i$), namely $\mathbf{F}_1$ and $\mathbf{F}_2$. By adding up these spectra in different proportions, we also obtained spectra of different mixtures $\mathbf{M}_i$ of these 2 computer-generated spectra. Below, some mixture spectra can be seen in Figure 5.7.



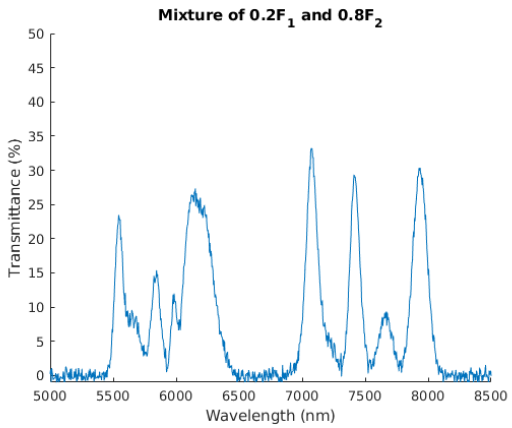
(a) Mixture 1 ($\mathbf{M}_3$).



(b) Mixture 2 ($\mathbf{M}_5$).



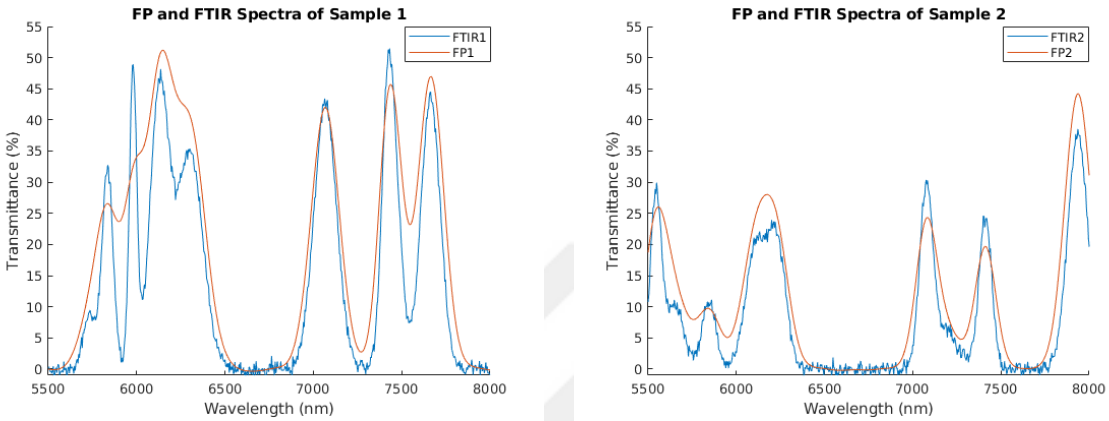
(c) Mixture 3 ($\mathbf{M}_7$).



(d) Mixture 4 ($\mathbf{M}_9$).

Figure 5.7: Some example mixtures of generated FTIR sample spectra. $\mathbf{M}_i$ represent mixture of two different samples $\mathbf{F}_1$ and $\mathbf{F}_2$ in different proportions.

Then, we created blurred spectra (called as FP) by multiplying the computer-generated FTIR spectra by blurring matrix $\mathbf{H}$, or convolution matrix, which was obtained by interpolating the given blurring kernels by the producer of the interferometers at only a few wavelengths. Some samples of blurred spectra can be seen in Figure 5.8.



(a) Sample 1. (b) Sample 2.

Figure 5.8: Generated ground truth FTIR spectra and blurred FP spectra for two different samples $\mathbf{F}_1$ and $\mathbf{F}_2$.

5.4.2. Reconstruction by LASSO

In reality, we had not FTIR spectra but FP spectra. Also, we could not truly interpret the properties of a sample by looking at its FP spectrum because it was overly smoothed version of FTIR, or true spectrum. Therefore, we needed to recover underlying FTIR spectrum from FP spectrum. To do this, reconstruction by L1-norm was used because we knew that there were lots of values close to zero in an FTIR spectrum. Firstly, our objective function for reconstruction may be seen below. Here, $\mathbf{z}$ is the underlying ground truth spectrum, which is FTIR spectrum in our case, that we want to reconstruct from FP spectrum, $\mathbf{H}$ is the convolution (blurring) matrix, $\mathbf{v}$ is the observed FP measurement and λ is the parameter for determining the sparsity of reconstructed spectrum.

$$f(\mathbf{z}) = ||\mathbf{v} - \mathbf{H}\mathbf{z}||_2^2 + \lambda||\mathbf{z}||_1 \tag{5.2}$$

The objective function above, which should be minimized, can be solved by iterative shrinkage algorithm. The proposed algorithm in the literature [41] is simply given as below. Here, the parameter c can take any value bigger than the maximum eigenvalue of $\mathbf{H}^T\mathbf{H}$. Also $S_{\lambda/c}$ is the soft thresholding operator with the threshold of λ/c.

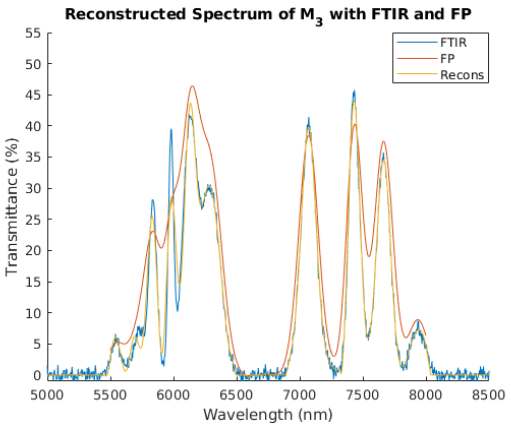
$$\mathbf{z}_{i+1} = S_{\lambda/c}\left(\frac{1}{c}\mathbf{H}^T(\mathbf{v} - \mathbf{H}\mathbf{z}_i) + \mathbf{z}_i\right) \tag{5.3}$$

By choosing $c = 1.1\lambda_{\max}(\mathbf{H}^T\mathbf{H})$ and $\lambda = 0.01c$, reconstruction was performed. Reconstructed spectra of samples and some of their mixtures with corresponding FTIR and FP spectra may be seen in Figure 5.9.

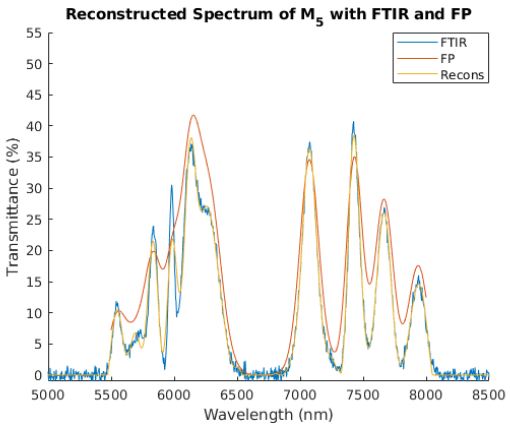
In order to assess the performance of reconstruction, the values at peak locations (wavelength) in FTIR spectra were plotted against the values at the same locations in corresponding reconstructed spectra for all mixtures. Plots of some arbitrary peaks may be seen in Figure 5.10. As can be seen, some peaks were almost perfectly reconstructed while a few of them had significant errors.

5.5. Discussion of Spectral Measurements

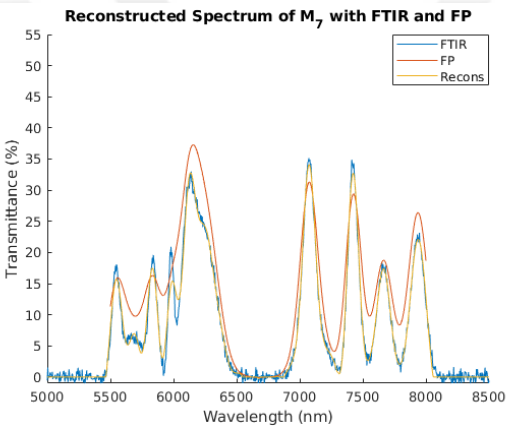
In this chapter, we proposed a potential spectroscopic setup with a commercial Fabry-Perot interferometer. Fast, portable and spectroscopic detection of some molecule footprints in blood serum data could pave the way for efficient and affordable prognostic solutions. However, every measurement setup come with its own limitation. One main drawback of the proposed setup is that measurements of the same sample at different times were significantly different. To solve this issue, we need to ensure that the sample dropped on the window to measure has always the same thickness. We processed a CaF_2 windows to keep the sample thickness constant but could not achieve a robust solution. Potential solutions will be discussed in the last chapter.



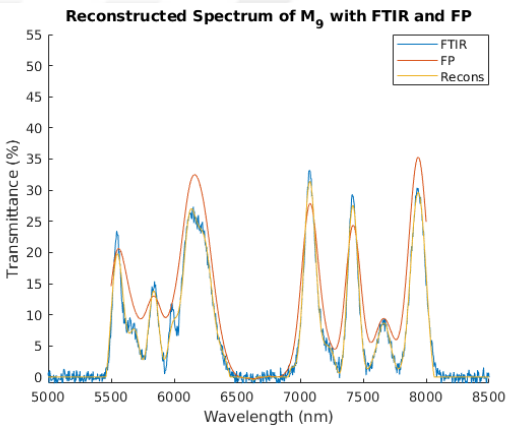
(a) Mixture 1 ($\mathbf{M}_3$).



(b) Mixture 2 ($\mathbf{M}_5$).



(c) Mixture 3 ($\mathbf{M}_7$).

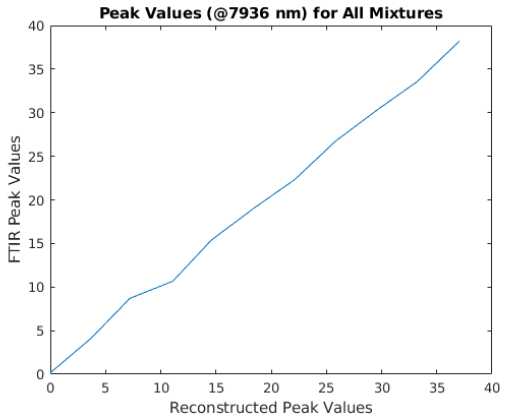


(d) Mixture 4 ($\mathbf{M}_9$).

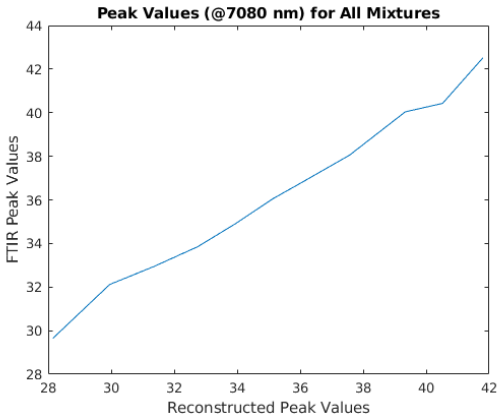
Figure 5.9: Reconstructed spectra of some example mixtures with their FP and FTIR spectra.

The other problem with the proposed measurement setup is the following. Because every interferometer has an inherent blurring kernel which corrupts the measurements, we tried to recover some computer-generated spectra with known kernel matrix. The results were fairly good because we could reconstruct FP spectra very similar to its ground truth, FTIR, by knowing the blurring kernel which caused smoothing.

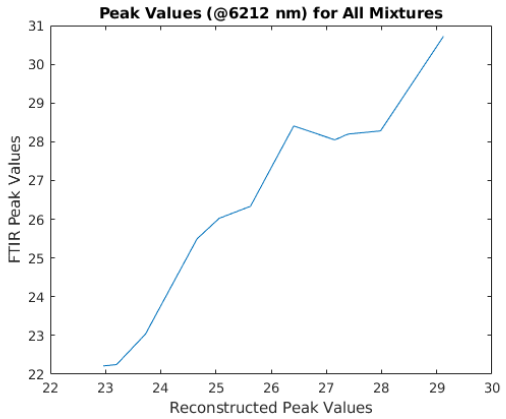
The main weakness of the spectral measurement setup is that we could not measure the performance of LASSO reconstruction on measured FP spectra because the blurring kernel matrix was not known for our FP device. We only know blurring kernel at a few wavelengths and interpolation is not a good solution when many points are not known. This created an ill-posed problem because we didn't know 2 out of 3 terms



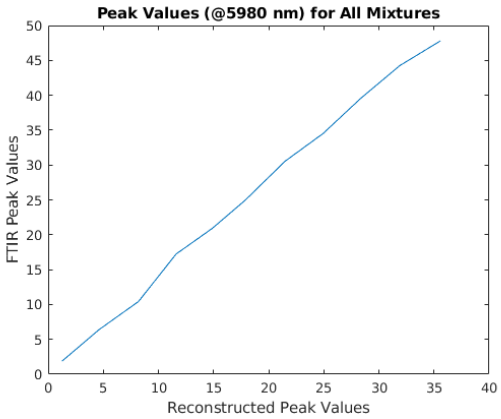
(a) Mixture 1 ($\mathbf{M}_3$).



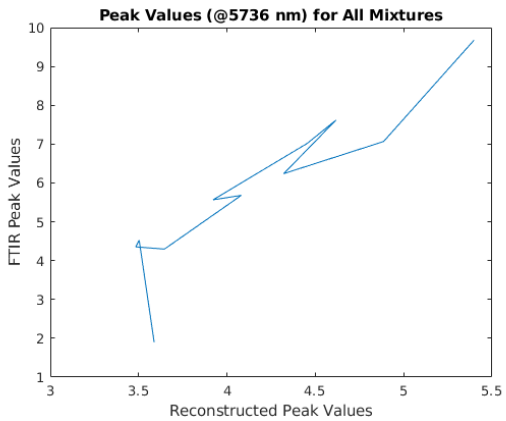
(b) Mixture 2 ($\mathbf{M}_5$).



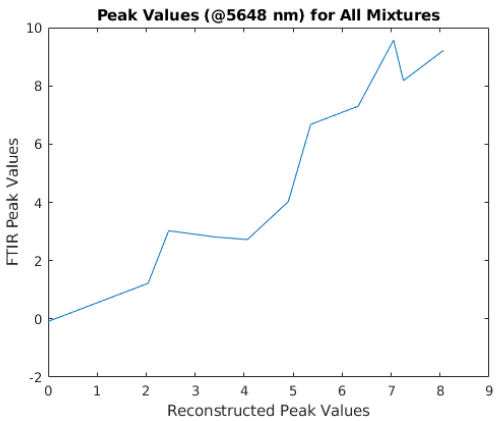
(c) Mixture 3 ($\mathbf{M}_7$).



(d) Mixture 4 ($\mathbf{M}_9$).



(e) Mixture 1 ($\mathbf{M}_3$).



(f) Mixture 1 ($\mathbf{M}_3$).

Figure 5.10: Comparison of peak values at different wavelengths in reconstructed and true FTIR spectra.

in classical inverse problem. Therefore, we cannot directly apply proposed LASSO solution on our FP measurements. Some potential solutions for these issues are argued in the last chapter.

6. CONCLUSIONS AND FUTURE STUDIES

For the first main pillar of this thesis, we identified biomarkers of breast cancer risk using metabolomics and protein profiling in plasma samples from healthy and susceptible individuals. We applied feature selection methods to potential biomarker candidates, then perform classification on the most promising ones using different training-test sets and report on the ability of these plasma biomarkers to discriminate between healthy and susceptible individuals. Future studies are required to validate these markers in bigger datasets.

For the other main pillar of this work, we proposed a novel FP spectroscopy based solution. Its drawbacks were introduced in Section 5.5. Robustness of measurements is the main thing that should be solved to make our device reliable. One possible solution may be constructing a micro-fluidic channel-like structure which would ensure a constant thickness of sample while light passes through it.

Another main problem which is introduced in Section 5.5 is the corruption of spectral measurements with a blurring kernel. A possible solution to this issue can be one-time characterization of blurring kernel at each wavelength by an expensive and more-advanced FTIR measurement device. Then, because iterative shrinkage algorithm gave good results when the blurring kernel was known, we can reconstruct true spectra of a sample from measured FP spectra signals. Another possible solution can be utilizing blind deconvolution methods. Potential benefits of using those techniques may be further investigated in future studies.

In conclusion, we first developed a pipeline to find new biomarkers from a huge collection of molecule measurements and then proposed a spectroscopic device which eventually can be utilized for fast, reliable and cheap measurement of those biomarkers. By easing the early detection of biomarkers for breast cancer, mortality rates especially in under-developed countries can be significantly reduced.

REFERENCES

1. Sun, Y. S., Z. Zhao, Z. N. Yang *et al.*, "Risk Factors and Preventions of Breast Cancer", *International Journal of Biological Sciences*, Vol. 13, pp. 1387–1397, 2017.

2. Ferlay, J., I. Soerjomataram, R. Dikshit *et al.*, "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012", *International journal of cancer*, Vol. 136, No. 5, pp. E359–E386, 2015.

3. Danaei, G., S. V. Hoorn, A. D. Lopez, C. J. Murray and M. Ezzati, "Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors", *Lancet*, Vol. 366, No. 9499, pp. 1784–1793, 2005.

4. Washbrook, E., "Risk factors and epidemiology of breast cancer", *Women's Health Medicine*, Vol. 3, No. 1, pp. 8–14, 2006.

5. Derks, M., E. Bastiaannet, W. van de Water *et al.*, "Impact of age on breast cancer mortality and competing causes of death at 10 years follow-up in the adjuvant TEAM trial", *European Journal of Cancer*, Vol. 99, pp. 1–8, 2018.

6. Coleman, M. P., M. Quaresma, F. Berrino *et al.*, "Cancer survival in five continents: a worldwide population-based study (CONCORD)", *The lancet oncology*, Vol. 9, No. 8, pp. 730–756, 2008.

7. Le Du, F., N. T. Ueno and A. M. Gonzalez-Angulo, "Breast cancer biomarkers: utility in clinical practice", *Current breast cancer reports*, Vol. 5, No. 4, pp. 284–292, 2013.

8. Duffy, M., N. Harbeck, M. Nap, R. Molina, A. Nicolini, E. Senkus and F. Cardoso, "Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM)", *European journal of cancer*, Vol. 75, pp. 284–298, 2017.

9. James, C. R., J. E. Quinn, P. B. Mullan, P. G. Johnston and D. P. Harkin, "BRCA1, a potential predictive biomarker in the treatment of breast cancer", *The oncologist*, Vol. 12, No. 2, pp. 142–150, 2007.

10. Walsh, M. F., K. L. Nathanson, F. J. Couch and K. Offit, "Genomic Biomarkers for Breast Cancer Risk", *Advances in experimental medicine and biology*, Vol. 882, pp. 1–32, 2016.

11. Nadal, R., A. Fernandez, P. Sanchez-Rovira *et al.*, "Biomarkers characterization of circulating tumour cells in breast cancer patients", *Breast Cancer Research*, Vol. 14, No. 3, p. R71, 2012.

12. Pernas, S. and S. M. Tolaney, "HER2-positive breast cancer: new therapeutic frontiers and overcoming resistance", *Therapeutic advances in medical oncology*, Vol. 11, p. 1758835919833519, 2019.

13. An, X., F. Xu, R. Luo *et al.*, "The prognostic significance of topoisomerase II alpha protein in early stage luminal breast cancer", *BMC cancer*, Vol. 18, No. 1, p. 331, 2018.

14. Ignatiadis, M., F. Rothe, C. Chaboteaux *et al.*, "HER2-positive circulating tumor cells in breast cancer", *PloS one*, Vol. 6, No. 1, p. e15624, 2011.

15. Cohen, J. D., A. Javed, C. Thoburn *et al.*, "Combined circulating tumor DNA and protein biomarker-based liquid biopsy for the earlier detection of pancreatic cancers", *Proceedings of the National Academy of Sciences of the United States of America*, p. 10202–10207, 2017.

16. Heywang-Köbrunner, S. H., A. Hacker and S. Sedlacek, "Advantages and disadvantages of mammography screening", *Breast care*, Vol. 6, No. 3, pp. 199–207, 2011.

17. Geisel, J., M. Raghu and R. Hooley, "The role of ultrasound in breast cancer

screening: the case for and against ultrasound", *Seminars in Ultrasound, CT and MRI*, Vol. 39, pp. 25–34, Elsevier, 2018.

18. Nothacker, M., V. Duda, M. Hahn *et al.*, "Early detection of breast cancer: benefits and risks of supplemental breast ultrasound in asymptomatic women with mammographically dense breast tissue. A systematic review", *BMC cancer*, Vol. 9, No. 1, p. 335, 2009.

19. Saslow, D., C. Boetes, W. Burke *et al.*, "American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography", *CA: a cancer journal for clinicians*, Vol. 57, No. 2, pp. 75–89, 2007.

20. Radhakrishna, S., S. Agarwal, P. M. Parikh *et al.*, "Role of magnetic resonance imaging in breast cancer management", *South Asian journal of cancer*, Vol. 7, No. 2, p. 69, 2018.

21. Heller, S. L. and L. Moy, "Breast MRI Screening: Benefits and Limitations", *Current Breast Cancer Reports*, Vol. 8, No. 4, pp. 248–257, 2016.

22. Kazarian, A., O. Blyuss, G. Metodieva *et al.*, "Testing breast cancer serum biomarkers for early detection and prognosis in pre-diagnosis samples", *British journal of cancer*, Vol. 116, No. 4, p. 501, 2017.

23. Hwa, H. L., W.-H. Kuo, L. Y. Chang, M. Y. Wang, T. H. Tung, K. J. Chang and F. J. Hsieh, "Prediction of breast cancer and lymph node metastatic status with tumour markers using logistic regression models", *Journal of evaluation in clinical practice*, Vol. 14, No. 2, pp. 275–280, 2008.

24. Santillán-Benítez, J. G., H. Mendieta-Zerón, L. M. Gómez-Oliván, J. J. Torres-Juárez, J. M. González-Bañales, L. V. Hernández-Peña and A. Ordóñez-Quiroz, "The Tetrad BMI, Leptin, Leptin/Adiponectin (L/A) Ratio and CA 15-3 are Reliable Biomarkers of Breast Cancer", *Journal of clinical laboratory analysis*, Vol. 27, No. 1, pp. 12–20, 2013.

25. Dalamaga, M., G. Sotiropoulos, K. Karmaniolas, N. Pelekanos, E. Papadavid and A. Lekka, "Serum resistin: a biomarker of breast cancer in postmenopausal women? Association with clinicopathological characteristics, tumor markers, inflammatory and metabolic parameters", *Clinical biochemistry*, Vol. 46, No. 7-8, pp. 584–590, 2013.

26. Assiri, A., H. F. Kamel and M. F. Hassanien, "Resistin, visfatin, adiponectin, and leptin: risk of breast cancer in pre-and postmenopausal saudi females and their possible diagnostic and predictive implications as novel biomarkers", *Disease markers*, Vol. 2015, 2015.

27. Provatopoulou, X., G. P. Georgiou, E. Kalogera *et al.*, "Serum irisin levels are lower in patients with breast cancer: association with disease diagnosis and tumor characteristics", *BMC cancer*, Vol. 15, No. 1, p. 898, 2015.

28. Patrício, M., J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seiça and F. Caramelo, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer", *BMC cancer*, Vol. 18, No. 1, p. 29, 2018.

29. Freitas, M., H. P. Nouws and C. Delerue-Matos, "Electrochemical Sensing Platforms for HER2-ECD Breast Cancer Biomarker Detection", *Electroanalysis*, Vol. 31, No. 1, pp. 121–128, 2019.

30. Tigli, O., L. Bivona, C. Chaterjee, M. E. Zaghloul and P. Berg, "Surface acoustic wave based biosensor in CMOS for cancer biomarker detection", *SENSORS, 2008 IEEE*, pp. 1452–1455, IEEE, 2008.

31. Pandya, H. J., K. Park, W. Chen, M. A. Chekmareva, D. J. Foran and J. P. Desai, "Simultaneous MEMS-based electro-mechanical phenotyping of breast cancer", *Lab on a Chip*, Vol. 15, No. 18, pp. 3695–3706, 2015.

32. Sun, D., Y. Ran and G. Wang, "Label-free detection of cancer biomarkers using an in-line taper fiber-optic interferometer and a fiber Bragg grating", *Sensors*, Vol. 17,

No. 11, p. 2559, 2017.

33. Misek, D. E. and E. H. Kim, "Protein biomarkers for the early detection of breast cancer", *International journal of proteomics*, Vol. 2011, 2011.

34. Gast, M.-C. W., E. J. van Dulken, T. K. van Loenen *et al.*, "Detection of breast cancer by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry tissue and serum protein profiling", *The International journal of biological markers*, Vol. 24, No. 3, pp. 130–141, 2009.

35. Alpaydin, E., *Introduction to machine learning*, MIT press, 2014.

36. Goossens, N., S. Nakagawa, X. Sun and Y. Hoshida, "Cancer biomarker discovery and validation", *Translational cancer research*, Vol. 4, No. 3, p. 256, 2015.

37. Hines, S. J., C. Organ, M. J. Kornstein and G. W. Krystal, "Coexpression of the c-kit and stem cell factor genes in breast carcinomas.", *Cell growth & differentiation: the molecular biology journal of the American Association for Cancer Research*, Vol. 6, No. 6, pp. 769–779, 1995.

38. Harrell, J. C., T. M. Shroka and B. M. Jacobsen, "Estrogen induces c-Kit and an aggressive phenotype in a model of invasive lobular breast cancer", *Oncogenesis*, Vol. 6, No. 11, p. 396, 2017.

39. Ebermann, M., K. Hiller, S. Kurth and N. Neumann, "Design, operation and performance of a Fabry-Perot-Based MWIR microspectrometer", *Sensor+ Test Conference*, Vol. 2009, 2009.

40. Thorlabs, *Optical Substrates*, 2019, `https://www.thorlabs.com/newgrouppage9.cfm?objectgroup_id=6973&tabname=Calcium%20Fluoride`, accessed at June 2019.

41. Daubechies, I., M. Defrise and C. De Mol, "An iterative thresholding algorithm for

linear inverse problems with a sparsity constraint", *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, Vol. 57, No. 11, pp. 1413–1457, 2004.