Old Dominion University

ODU Digital Commons

Combining Molecular and Imaging Biomarkers to Enhance MALDI Biomarker Analysis

Ayyappa Chowdary Vadlamudi
Old Dominion University

Recommended Citation

COMBINING MOLECULAR AND IMAGING BIOMARKERS TO

ENHANCE MALDI BIOMARKER ANALYSIS

by

Ayyappa Chowdary Vadlamudi
B.Tech. April 2008, Acharya Nagarjuna University, India

A Thesis Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

ELECTRICAL ENGINEERING

OLD DOMINION UNIVERSITY
May 2010

Approved by:

Jiang Li (Director)

Yuzhong Shen (Member)

Frederic McKenzie (Member)

ABSTRACT

COMBINING MOLECULAR AND IMAGING BIOMARKERS TO ENHANCE MALDI
BIOMARKER ANALYSIS

Ayyappa Chowdary Vadlamudi
Old Dominion University, May 2010
Director: Dr. Jiang Li

This thesis presents a three-step method to predict prostate cancer (PCa) regions on biopsy tissue samples based on high confident, low resolution PCa regions marked by a pathologist. First, a prediction model is designed to predict PCa regions using matrix-assisted laser desorption mass spectrometry (MALDI-MS) tissue imaging data from one prostate tissue slice. Second, a texture analysis technique is applied to a high magnification optical image for the same purpose from an adjacent tissue slice. Finally, those two results are fused to obtain the PCa regions that will assist MALDI imaging biomarker analysis. Experiments show that the texture analysis based prediction is sensitive but not specific, and the prediction based on the MALDI-MS data is specific while less sensitive. By combing those two results, a much better prediction for PCa regions on the adjacent slice can be achieved. This thesis focuses on the MALDI-MS based PCa region prediction and fusion of prediction from texture analysis and that from MALDI-MS data processing.

ACKNOWLEDGMENTS

I would like to express my sincere thanks to Dr. Jiang Li for his support, advice, and motivation during my research and study at the Old Dominion University. Without his constant guidance, it would not have been possible to complete this thesis.

I would like to thank Dr. Frederic McKenzie and Dr. Yuzhong Shen for agreeing to be on my thesis committee and for their valuable time. I want to thank all members of my research group for their valuable suggestions.

I would like to thank Professor Dr. John Semmes's group at the East Virginia Medical School (EVMS) for providing the mass spectrum data and prostate tissue used in this thesis.

Finally, I would like to thank my family and all my friends for being there for me all the time.

This thesis is dedicated to my parents.

TABLE OF CONTENTS

LIST OF FIGURES

Figure   Page

LIST OF TABLES

CHAPTER 1

INTRODUCTION

This chapter gives a brief introduction to prostate cancer and its effects. The methods involved in the diagnosis of prostate cancer are discussed. The goal of this thesis and the work proposed are presented in this chapter.

1.1 Background

Prostate cancer is a disease that develops in the prostate gland, a male reproductive element in the size of a walnut and present beneath the bladder just opposite the rectum. Prostate cancer is a dreaded non-skin cancer that kills one man every nineteen minutes, and studies show that every one in six men will suffer from this disease [1]. Prostate cancer is the most common type of cancer among American men, killing 186,000 individuals a year [2]. The rate of prostate cancer is not uniform over the world, but it varies across nations. It is most common in United States and more common in the European countries; it becomes quite less common in the Asian countries [3]. Prostate cancer is difficult to cure if it metastasizes to other organs, so early detection is demanded. Currently, the prostate specific antigen (PSA) test is the most common early diagnosis method, where a blood sample is taken and the level of PSA is checked [4]. A high level (approx 10ng/ml) is considered positive for prostate cancer. However, recent studies prove that some patients with low PSA concentrations are also affected with prostate cancer. Fifteen percent of men have prostate cancer even if their PSA concentrations are low [5]. This shows that the PSA test is prostate specific rather than prostate cancer specific. The benign prostatic hyperplasia is observed to contain more increased concentrations of PSA [6], which

---

The reference model for this work is *IEEE Transactions on Image Processing*

confronts the above statement. This shows that the PSA level is not an effective tool for early detection of prostate cancer. Therefore, there is a need to discover more accurate and specific detection methods for prostate cancer.

1.2 Biomarkers and Mass Spectrometry

Biomarkers refer to specific genes and their products (proteins) that are said to be the indicators of the diseased states [7]. Biomarkers must be accurate, sensitive and specific, and be able to differentiate between the normal and diseased populations. With the development of technology, protein mass spectrometry has evolved to be a useful tool to identify molecular biomarkers. The two most widely used methods involve surface-enhanced laser desorption/ionization (SELDI) and matrix-assisted laser desorption/ionization (MALDI) time of flight (TOF) approaches [8]. Out of these two methods, MALDI-TOF proves to be more amenable to a higher throughput analysis. A mass spectrum can be represented as a vector whose dimensionality is equal to the number of distinct m/z values recorded by the spectrometer, and the value of each dimension is the intensity of the molecule with the corresponding m/z value. Simply, it is a curve with m/z values on the x-axis and signal intensities on the y-axis. Every peak corresponds to a specific protein/molecule, and our goal is to identify peaks that are related to specific outcomes or malignancy stages.

1.3 Histopathological Examination

Cancer tissue regions are usually identified by pathologists after examining hematoxylin & eosin (H&E) stained histological microscopy images. Unfortunately, histopathological examination is currently done on an adjacent slice because the H&E staining process will change

the tissue's protein structure, and it will derogate MALDI analysis if the same tissue is used. At the same time, the MALDI imaging process will destroy the tissue slice so that it is no longer available for histopathological exam. For this reason, only the most confident cancer region resulting from the histopathological examination on an adjacent slice will be used to guide the biomarker identification.

1.4 Challenges Faced

The first and foremost challenge is the destruction of the tissue slice by the MALDI imaging process. For this reason, the region that is marked by a pathologist on an adjacent tissue slice is taken as the reference for biomarker identification. Another challenge is how to effectively select peaks (features) that can discriminate cancer spectra from normal ones. There are many feature selection algorithms used for the biomarker identification. Lyons-Weilera et al. discussed many feature selection algorithms like the AUC score, J5 test, mRMR, Random search and genetic algorithm search methods that are utilized for the biomarker identification [9]. In the AUC score method, a receiver operator characteristics (ROC) curve is computed for each of the features and is integrated over the curve to get an area under the ROC curve (AUC) score. A higher AUC score implies a better feature [10]. The J5 test is the ratio of the intensity difference of two groups to that of the mean difference of all genes in a group. The method of selecting features that are distinct from each other but have a high correlation with the target variable is called the Minimum-Redundancy-Maximum-Relevance selection [11]. The feature selection method in which random subsets of features are generated and their quality is assessed independently by a learning algorithm, which further selects the best set of features, is called a random search algorithm [12]. Feature selection for biomarker identification has been tackled

previously by our group [13, 14]. This thesis focuses on the first challenge, that is, to predict an accurate cancer boundary on the adjacent tissue slice by combining a texture analysis result and the result from a MALDI MS process.

1.5 Goal

The goal of this thesis is to provide a better estimate of the cancer boundary through the combination of MALDI MS processing and texture analysis on microscopy images.

1.6 Proposed Work

In this thesis, a three-step method is presented to predict prostate cancer (PCa) regions on biopsy tissue samples based on high confidence, low resolution PCa regions marked by a pathologist. First, a prediction model is designed using matrix-assisted laser desorption mass spectrometry (MALDI-MS) tissue imaging data to predict the true PCa regions. Second, a texture analysis technique is applied on a high magnification optical image from an adjacent slice to predict true PCa regions on the slice subjected to MALDI imaging. Finally, those two results were fused to obtain the PCa regions on the slice subjected to MALDI imaging, which assists the MALDI biomarker identification.

1.7 Thesis Outline

Chapter 2 presents the methods of acquiring the data, the software used and brief explanations of the methods used in the thesis, such as the MALDI imaging technique and the texture analysis method. It also presents a registration method used in the thesis.

Chapters 3 and 4 give detailed information on the steps to preprocess the mass spectra

data such as baseline adjustment, smoothing, normalization, peak detection, clustering and classification. Preprocessing of optical images for texture analysis will be introduced as well.

Chapter 5 provides performance metrics and tools used in our experiments such as the k-fold cross validation method, the sensitivity and the specificity measurements, and the confusion matrix table.

Chapter 6 shows our experimental designs for MALDI-MS data processing and texture analysis. This chapter also gives the results of the experiments conducted.

Chapter 7 describes the fusion method of combing the results from the MALDI-MS data processing and texture analysis to achieve a better estimation model, and chapter 8 concludes the work.

CHAPTER 2

METHODS

A prostate biopsy tissue was obtained from the Eastern Virginia Medical School (EVMS) from which two adjacent slices were taken. One tissue slice was used to obtain MALDI MS data and another H&E stained slice was utilized for pathological examination and for texture analysis.

2.1 The Specimen

The tissue obtained from the EVMS is shown in the Fig. 2.1. The cancer region on the tissue is circled and a close-up of the cancer region is shown on the right.

2.2 BioMap Software

The BioMap software [15] is a visualization tool that shows the spatial distribution of proteins across the tissue. The software takes the MALDI MS data obtained from the whole tissue slice (adjacent to the tissue slice shown in Fig. 2.1) as inputs. It then shows the spatial distribution of a specified protein (or m/z value) across the tissue. Fig. 2.2 shows the spatial distribution of a biomarker identified at EVMS. It is clear that the biomarker can roughly differentiate the cancer region from normal regions, as the cancer region has a high concentration of the identified biomarker as shown in red.
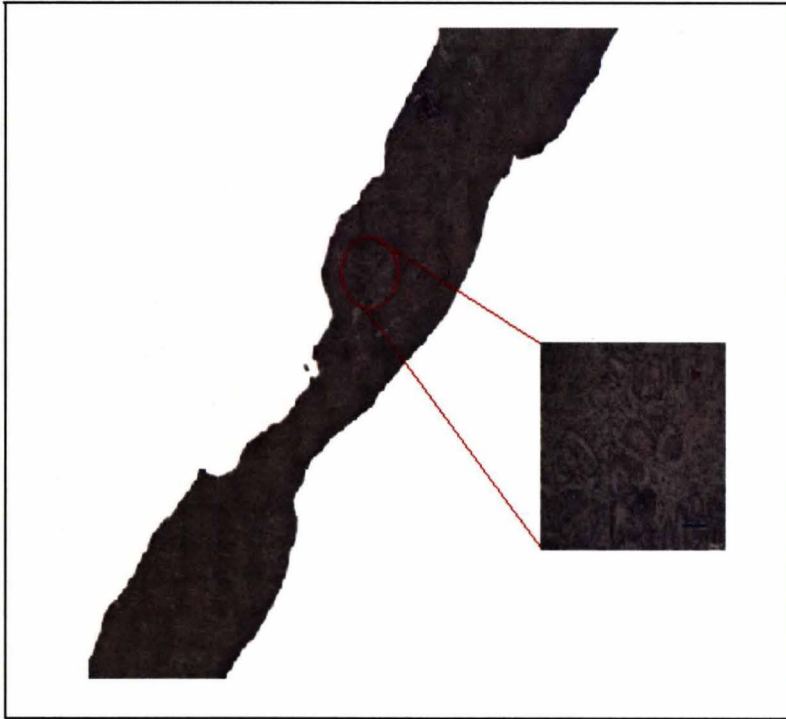
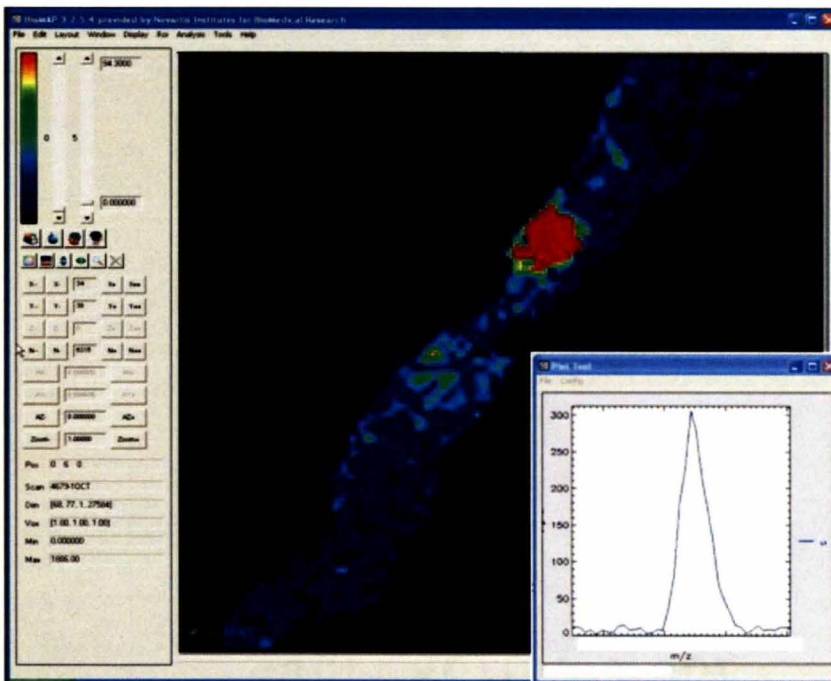**Fig. 2.1** The specimen tissue used


Fig. 2.2 BioMap software

2.3 MALDI Imaging

MALDI imaging is a soft ionization technique used in mass spectrometry for the analysis of the bio-molecules such as peptides and proteins. A beam of laser hits the tissue slice at spots with a size of 100 micrometers in diameter across the entire tissue, yielding a MS spectrum from each of the 974 spots on the slice. As MALDI is a time-of-flight approach, a mass spectrum is represented by a vector whose dimensionality is equal to the number of distinct m/z values recorded and the value of each dimension is the concentration of molecule with the corresponding m/z value. Fig. 2.3a and Fig. 2.3b show one normal spectrum and one cancer spectrum, respectively. The corresponding normal and cancer regions can be seen in the Fig. 2.3c and Fig. 2.3d. The co-ordinates of the spots were then used to construct an artificial image representing the shape of the tissue (Fig. 2.3e). This is called the MALDI co-ordinates visualization (MCV). The tumor part can be clearly seen in the image (white region in Fig. 2.3e, named AI1). This image was used as the low magnification ground truth (LMGT).

2.4 Histopathological Analysis

The H&E stained slice was scanned using a Hirox HI-SCOPE KH-1300 microscope at a magnification of 420x, resulting in a set of images. The H&E stained slice was captured in a total of 184 images (as shown in Fig. 2.3c and Fig. 2.3d). The whole tissue image was then reconstructed with at least 20% overlap with each other as shown in Fig. 2.4a, using the E-tiling software. The reconstructed image was observed and analyzed by a pathologist, on which the high confidence cancer regions were annotated. The pathologist classified cancer regions on the tissue image and these are shown in Fig. 2.4b. A mask image was produced for the H&E tissue slice along with the annotated cancer regions using Adobe Photoshop CS3 and was termed AI2.

The mask image can be seen in Fig. 2.4c. This image was used as high magnification ground truth (HMGT).
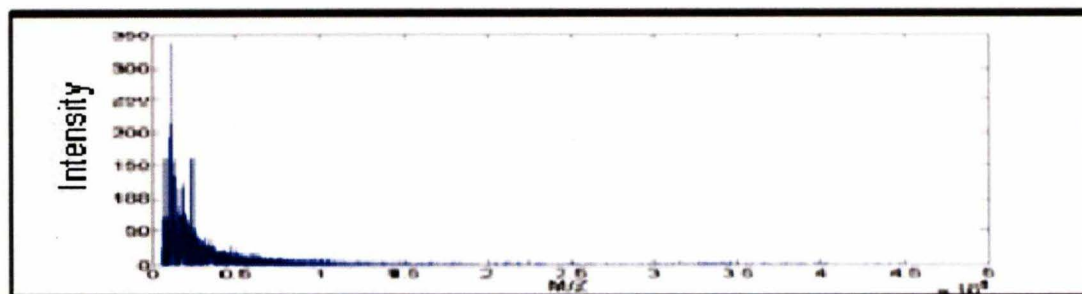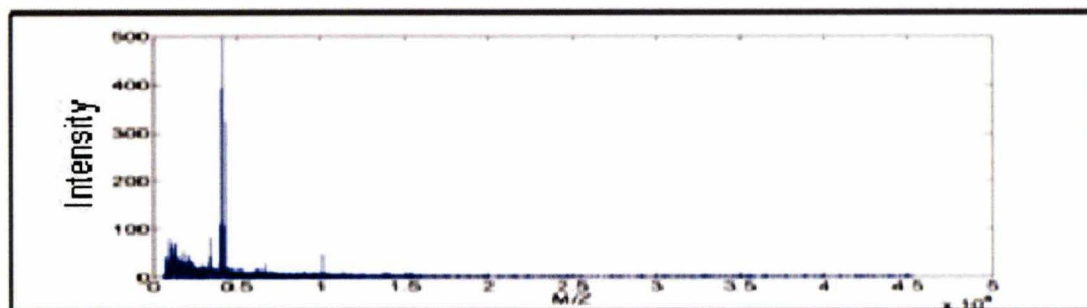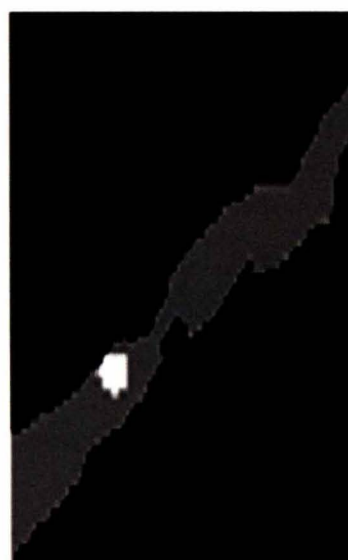
Fig. 2.3 a) Normal spectra, b) Cancer spectra, c) Normal region, d) Cancer region and e) MCV image
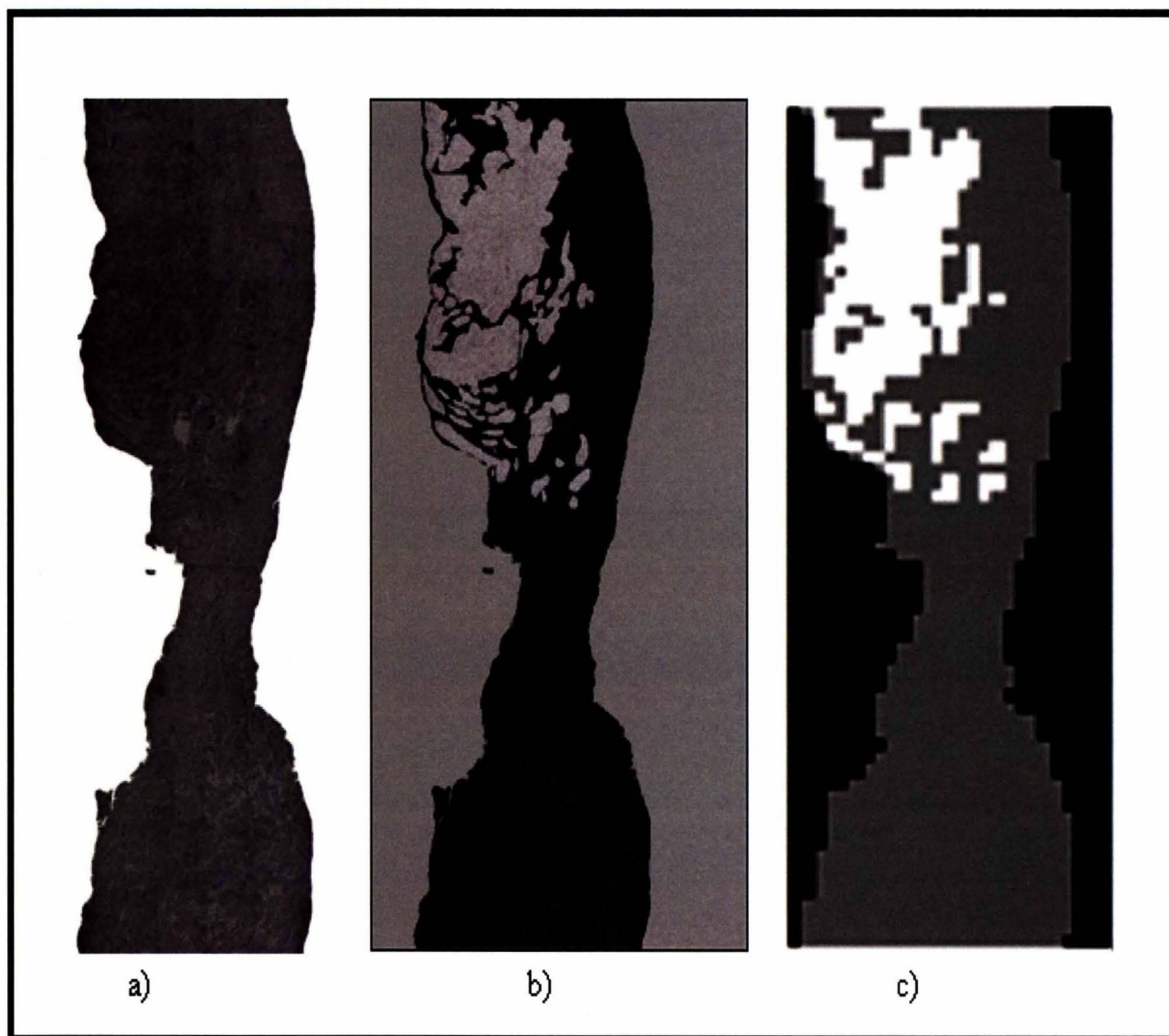
Fig. 2.4 a) Stitched image, b) Annotated regions and c) Mask image of the H&E tissue

2.5 Registration

The images obtained from the MALDI co-ordinates visualization and histopathological examinations are very much different from each other in size and orientation. For any further studies to be done on these images, they must be of the same size and orientation. Therefore, it is necessary to adopt a registration process. In this thesis, a landmark based registration method is utilized and implemented by MATLAB. In this type of registration, the input image is used to compare or align with the base or reference image. A control point selection tool is launched after the base and input images are read. A set of control points (landmarks) were selected from both the images and were passed to a function, which determined a transformation matrix based on the geometrical relationship of the control points. There are various transformation matrix types such as piecewise linear, affine, projective, etc. We can choose the transformation matrix type based on our need. In this thesis, affine transformation is used. The transformation function created a spatial transformation structure and then was applied to the input image. The steps followed along with the functions used can be seen in Fig. 2.5. An example for landmark registration and control points selection on the base image and input image using MATLAB built-in function is shown in Fig. 2.6. The above landmark registration was applied for registering the image from the histopathological examination (Fig. 2.4c, AI2) to the MALDI co-ordinates visualized image (Fig. 2.3e, AI1), for the MALDI process. In this process, AI1 is considered as the base image and AI2 as the input image, and the ground truth on AI2 was transferred to AI1 for a better MALDI analysis. The transformed image, which is the result from the registration process, can be seen in Fig. 2.7c. To confirm an exact match between the two images, AI2 is overlaid on AI1 (Fig. 2.7d). Small offsets can be observed as they are not the same tissue slice, but adjacent tissue slices.
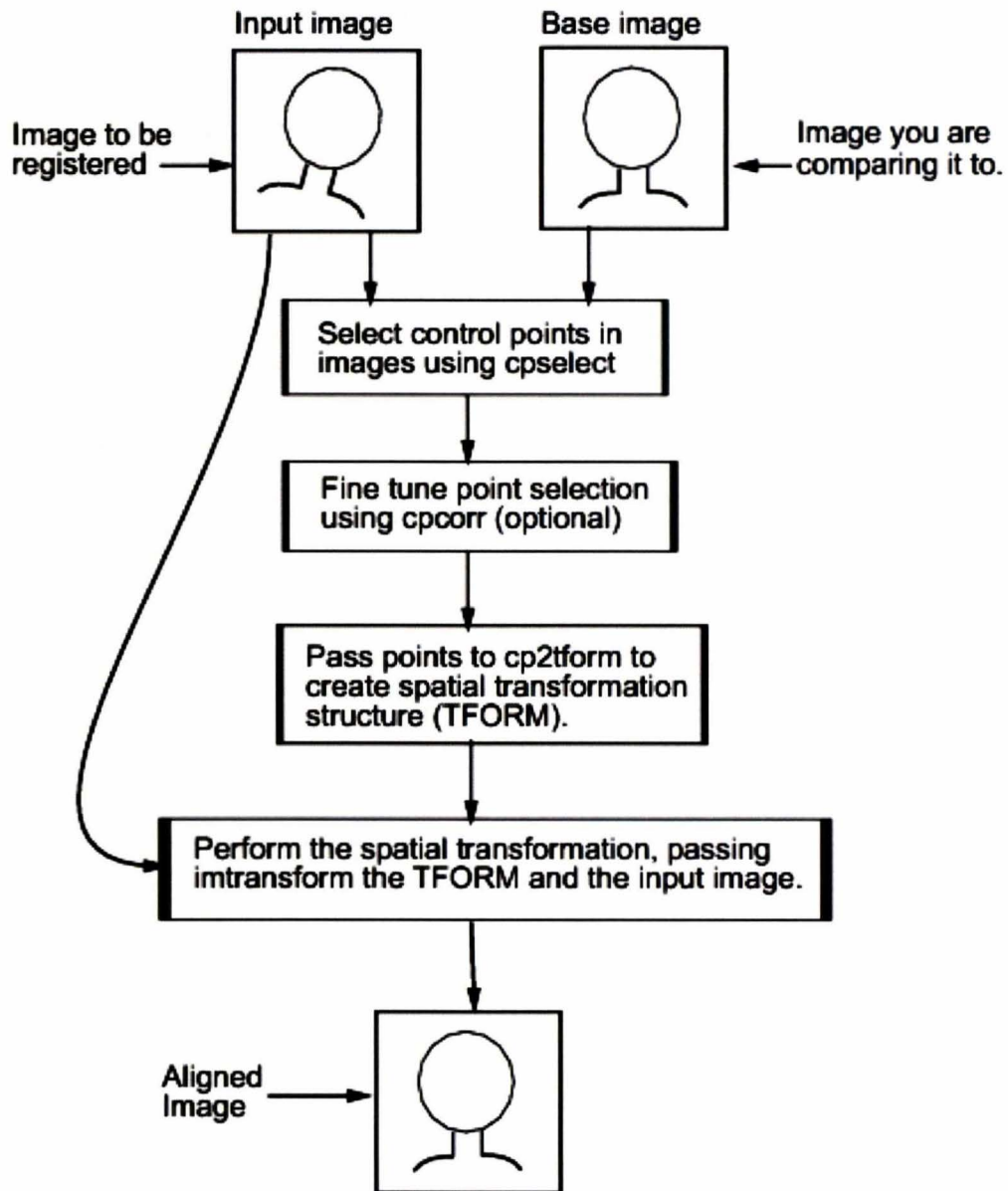
Fig. 2.5 Steps followed for landmark registration [16]
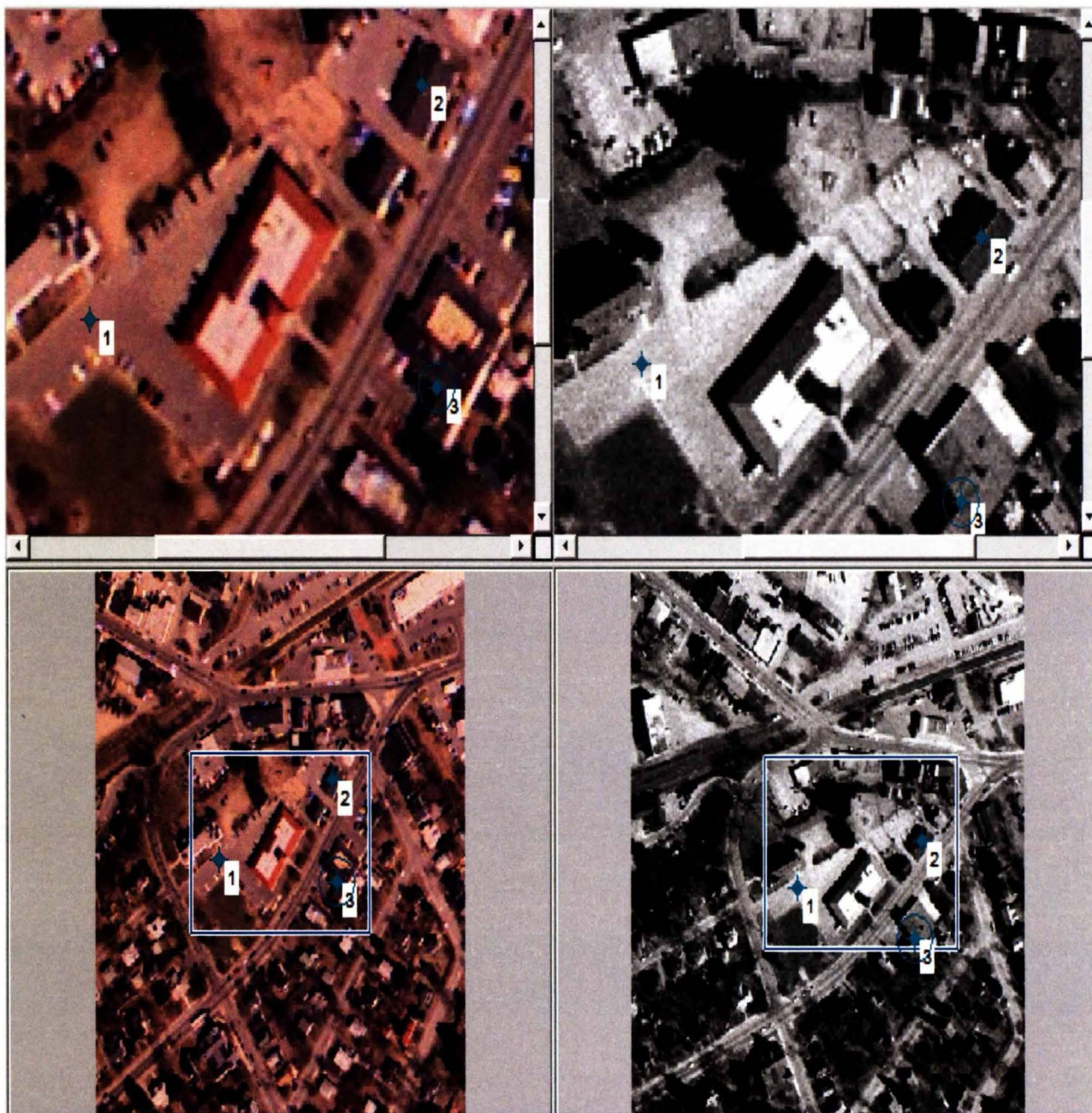
Fig. 2.6 Example of landmark registration. Courtesy of MATLAB.

The white part (tumor part) from the registered AI2 is mapped on to the MALDI co-ordinates visualized image (Fig. 2.7e). This is considered an improved ground truth in the

MALDI MS data space and was used for the MALDI MS method.

A similar registration approach was used for the texture analysis method. The histopathology image (Fig. 2.8a. AI1) of the tissue was used as the base image and MALDI co-ordinates visualized image was taken as the input image. The selected control points (landmarks) can be seen in Fig. 2.8b. After the image was registered, the new image was overlaid on the reference image to check for offsets (Fig. 2.8c.). As they are not of the same slice, but adjacent slices they do have some irregularities among them. The registered image can be seen in Fig. 2.8d. This registered image was used as the low magnification ground truth (LMGT) for the texture analysis method. In this thesis, as two methods are being used: the MALDI MS analysis and the texture analysis, registration was also done in two ways. Consequently, there are two ground truths in both the MALDI MS data and optical image data spaces: high magnification and low   magnification   ground   truths.
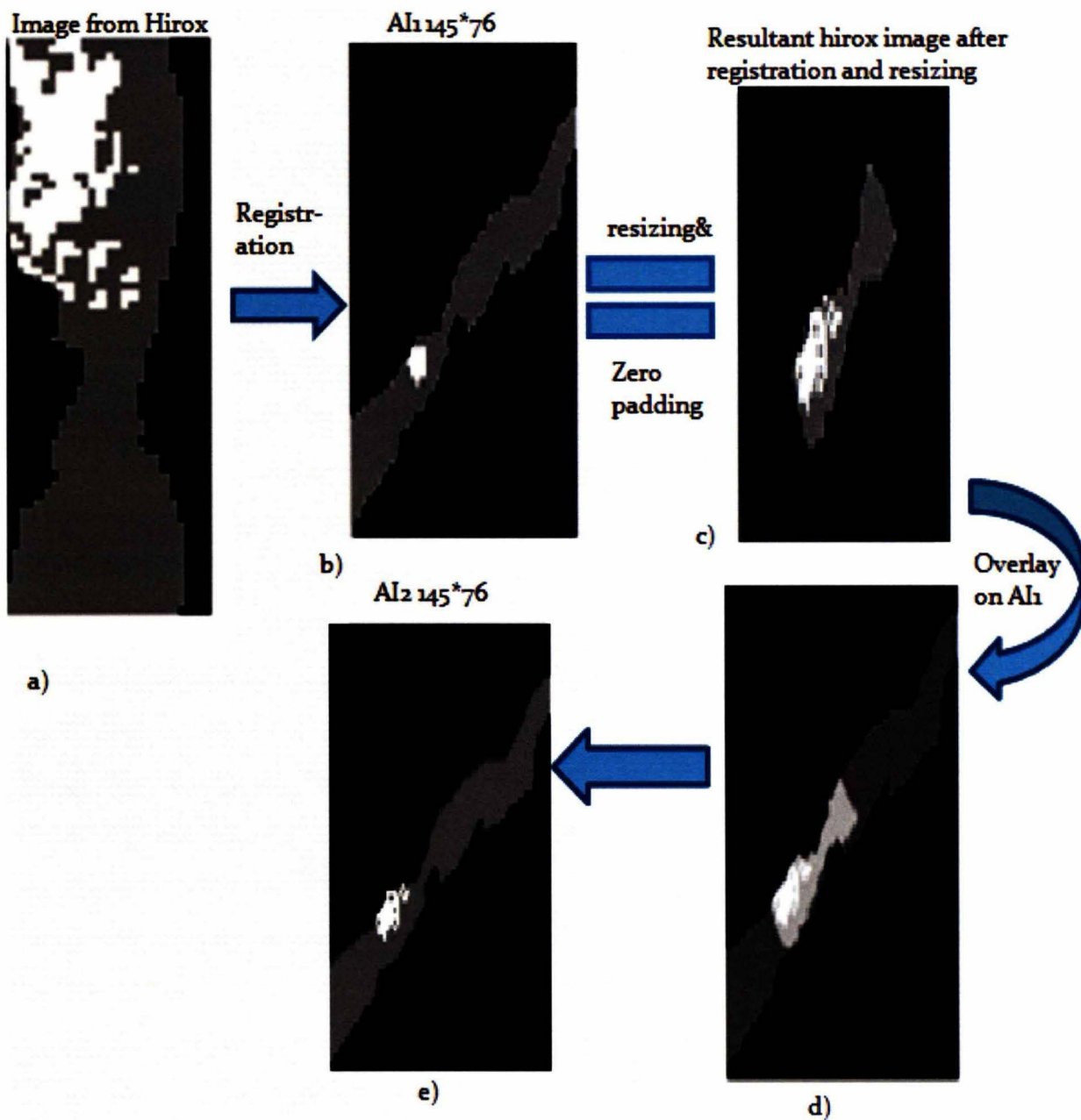
Fig. 2.7 Registration process of HMGT to LMGT: a) High resolution cancer region mask produced by the pathologist, b) Artificial tissue image constructed from the coordinates of the MALDI MS data, c) Registered high resolution image d) Overlaid ground truth AI1 and AI2 and e) Registered AI2 to MALDI MS data.
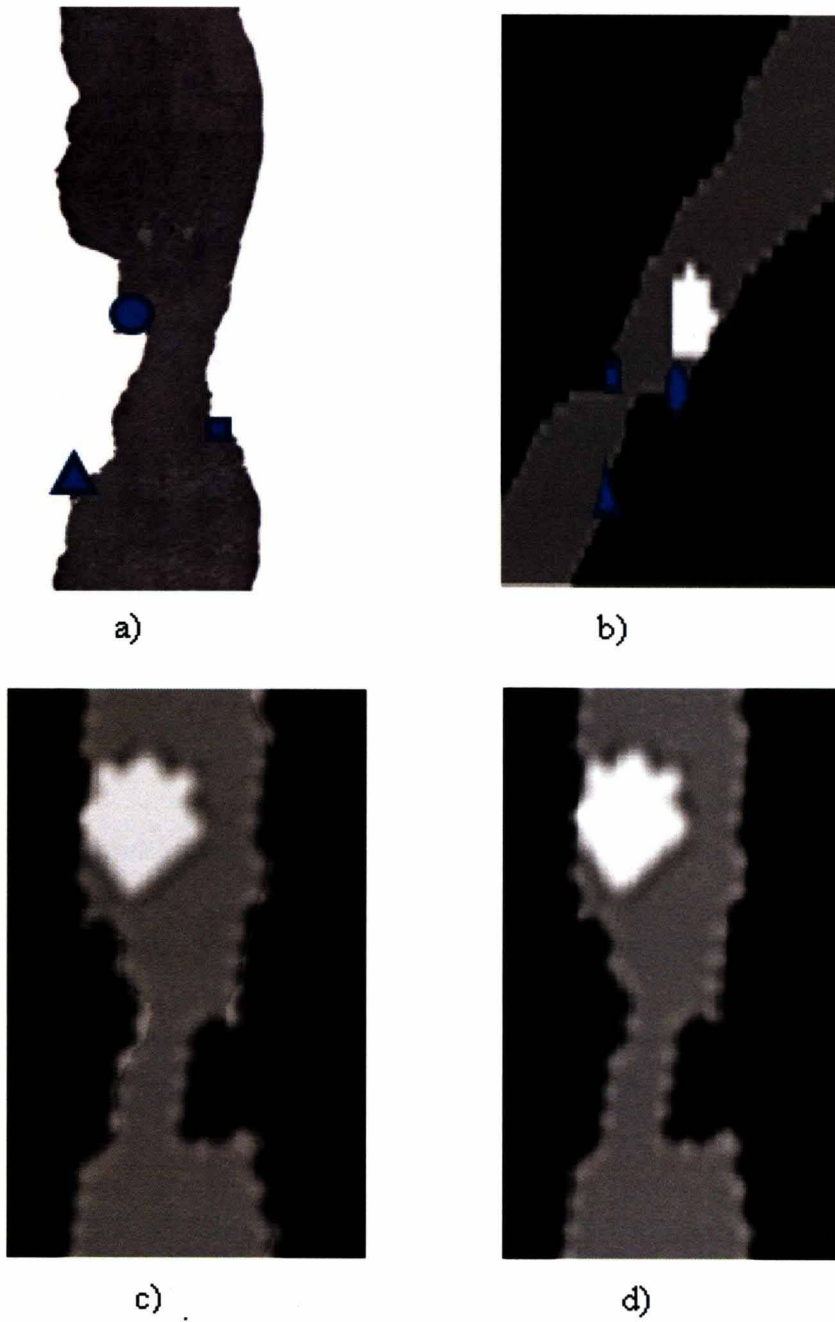
Fig. 2.8 Registration process of LMGT to HMGT: a) H&E image with landmarks, b) AI1 with landmarks, c) Overlay of images and d) New LMGT.

2.6 Ground Truths

There are three ground truths in this study for both MALDI MS data analysis and texture analysis, respectively: AI1, AI2 and AI1&AI2 (intersection of AI1 and AI2) as shown in the Fig. 2.9. For a better understanding, the AI1 was used as LMGT, AI2 was used as HMGT and the intersection was used as LMGT&HMGT. The ground truths used in both texture analysis and MALDI MS were the same except for different orientation and size due to the variation in registration methods chosen in the earlier section.
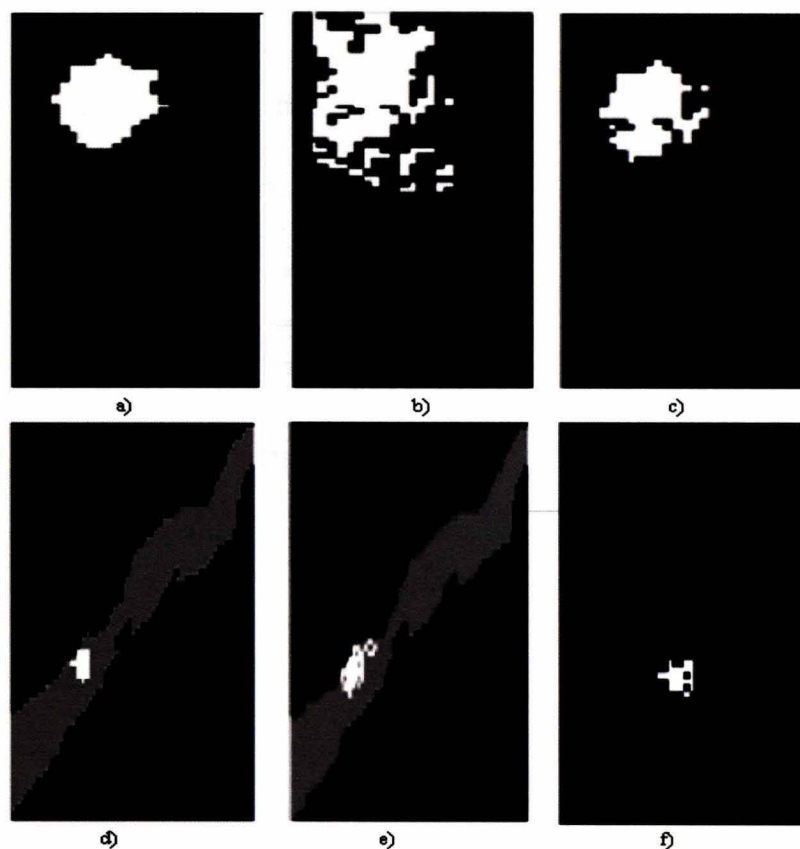
Fig. 2.9 Texture analysis: a) AI1 (LMGT), b) AI2 (HMGT) and c) AI1&AI2 (LMGT &HMGT). MALDI MS: d) AI1 (LMGT), e) AI2 (HMGT) and f) AI1&AI2 (LMGT &HMGT).

CHAPTER 3

MALDI IMAGING

This chapter discusses in detail the pre-processing techniques used, feature selection algorithm and the classification algorithm used in the MALDI imaging method.

3.1 Preprocessing Techniques

The raw MALDI data received from EVMS contains much unwanted noise, and therefore is needed to undergo a series of pre-preprocessing steps [13], which are described as follows.

3.1.1 Baseline adjustment

In general, it is recommended to remove the ion and chemical noise that are usually higher at smaller m/z values. As the baseline characteristics vary from one experiment to another, there exists no general solution to this problem. In this thesis, a Matlab function called msbackadj was used [14]. It estimates the baseline within multiple shifted windows and regresses the varying baseline using a spline approximation to the window points with a size of 200 m/z values. The results can be clearly seen in Fig. 3.1a. The regressed baseline is shown in red while the blue lines represent the original spectra. The black marks show the estimated baseline points. These can be clearly observed in the close-up Fig. 3.1b.

3.1.2 Smoothing

A wavelet based algorithm was used for denoising and enhancing the signal to noise ratio for the MALDI MS data. The m/z values that are lower than 3000 have more noise, and the m/z values greater than 10,000 have low intensities. Therefore, this particular set of intensities were discarded, as shown in the Fig. 3.2a.
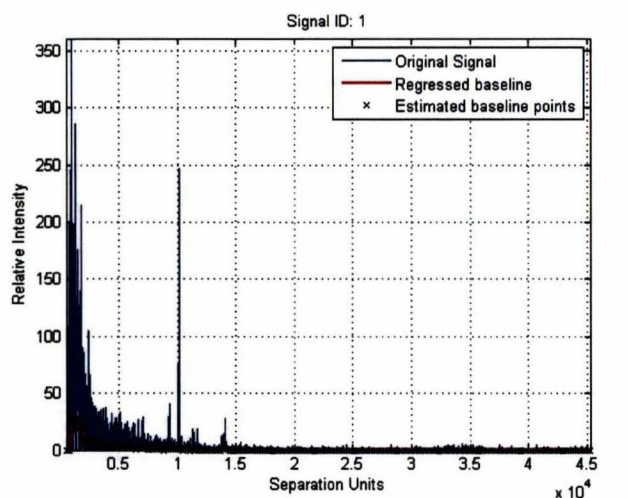
a)
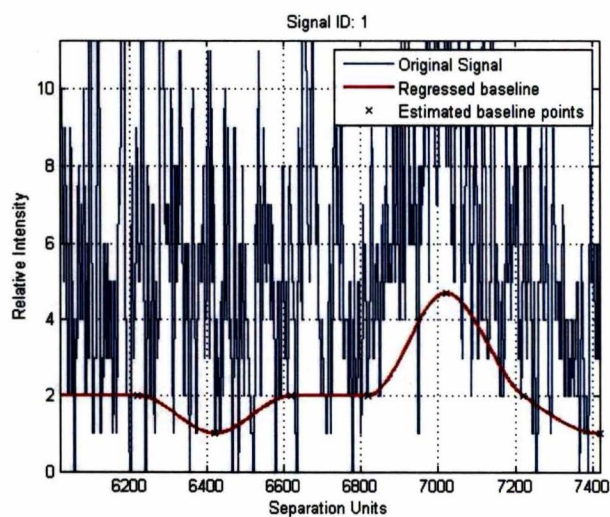
b)

Fig. 3.1 a) Result of baseline adjustment and b) close-up of Fig. 3.1a.

3.1.3 Normalization

With multiple spectra, problems such as systematic variation among spectra arise due to the detector sensitivity variation. It is a good practice to remove these types of effects by normalization. A common factor was used to scale the mass intensities for the same peak from different spectra [17]. For a given peak, the area under the peak was computed, the ratio of the area under this peak to the median of areas of all other peaks in a single spectrum was taken as the common factor.

3.1.4 Peak Detection

In order to identify and quantify the proteins in mass spectra, the crucial step is to find the m/z values that correspond to higher intensities [18]. A mass of points having good S/N ratio and reasonable intensities were selected. This method of peak detection satisfies the criteria that the intensity exceeds a specific threshold of 10, under which all other intensities are zeroed. One example of spectra with detected peaks is shown in Fig. 3.2a. After the process of smoothing and peak detection, a total of 75,719 peaks are obtained from the total 974 spectra. Each spot on the prostate tissue can be visualized as an individual spectrum. Fig. 3.2b shows the results of the smoothing, normalization and peak detection. The blue and green color in the spectrum shows the original and the denoised spectrum. The detected peaks are shown by the red markings. The unnecessary peaks and other typical noises were greatly reduced by the above pre-processing steps, which simplify the biomarker identification.
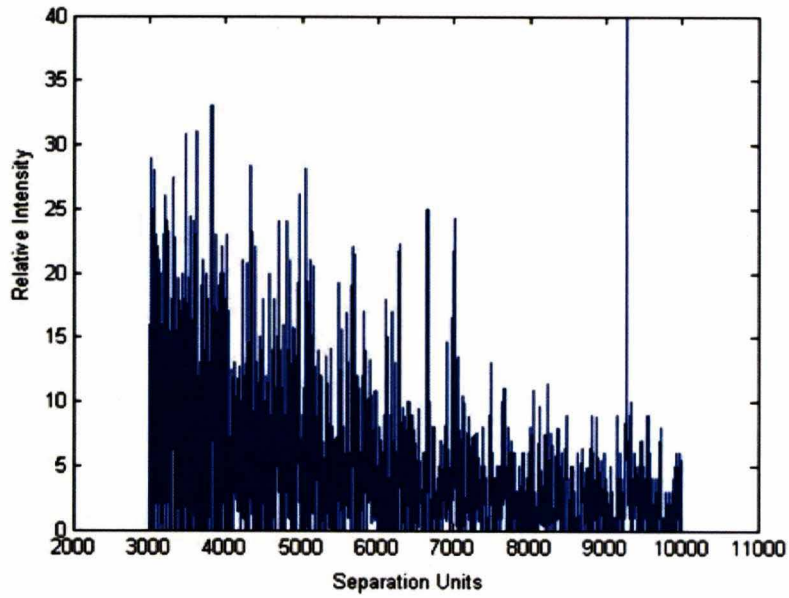
a)
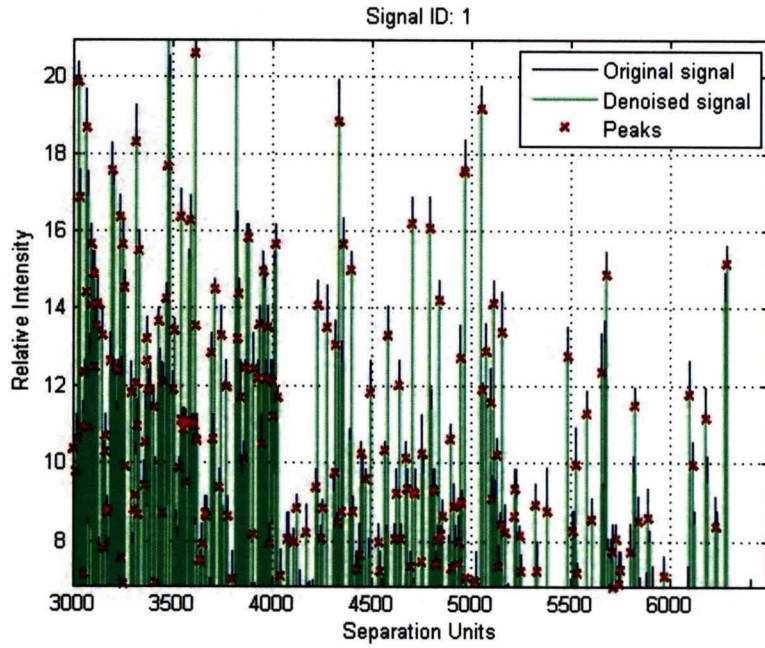
b)

Fig. 3.2 a) Result of smoothing and b) Result of normalization, smoothing and peak detection.

3.1.5 Clustering/re-binning

This step is to align or cluster the same peak from different spectra and to assign a cluster number to every peak found in all spectra with slightly different m/z values. All the detected peaks (from the 974 spectra) were projected on to a single axis and then clustered [19]. This can be seen in Fig. 3.3a and its close-up version in Fig. 3.3b. The same peak may have slightly different m/z values due to the fact that the spectra exhibits shifts in the horizontal axis. In this step, all the peaks that have m/z values within a variation of 0.13% of each other are merged and assigned the mean m/z value for all peaks in the cluster. This step yielded a total of 820 clusters, which represented 820 different peaks as shown in Fig. 3.4a and its zoom-in version in Fig. 3.4b. The red lines in the Figure show the 820 clusters. The next step is to back project these peaks onto individual spectra for identifying the biomarkers. If the individual spectrum consists of a peak that corresponds to a cluster, the peak's intensity was maintained the same. If the spectrum contains no such peak, then the intensity for the peak was replaced by zero. These results are shown in Fig. 3.5a and its zoom-in version in Fig. 3.5b. After all the raw data were preprocessed, a set of peaks were obtained and were denoted as $\{x_p, i_p\}_{p=1}^{Nv}$, where $x_p \in R^N$ and $i_p \in R$, x_p corresponds to a vector consisting of peaks detected from p^{th} spectrum and i_p is the class ID (1: normal, 2: cancer) associated with this spectrum. The total number of peaks from each of the spectra is N (820 in our study) and the total number of spectra is N_v (974 in this paper). The results from the pathological analysis determine the class ID and were considered as the ground truth. Note that we have different resolution ground truths so that a spectra might have a different class ID based on the considered ground truth.
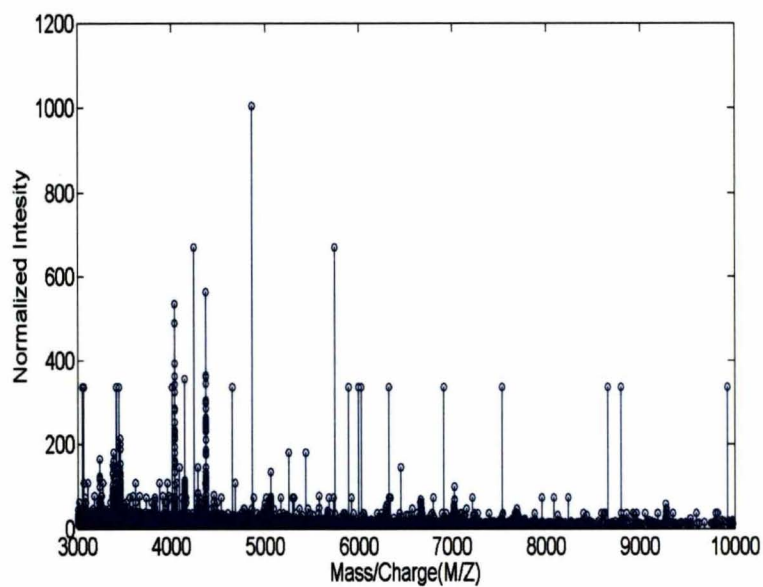
a)
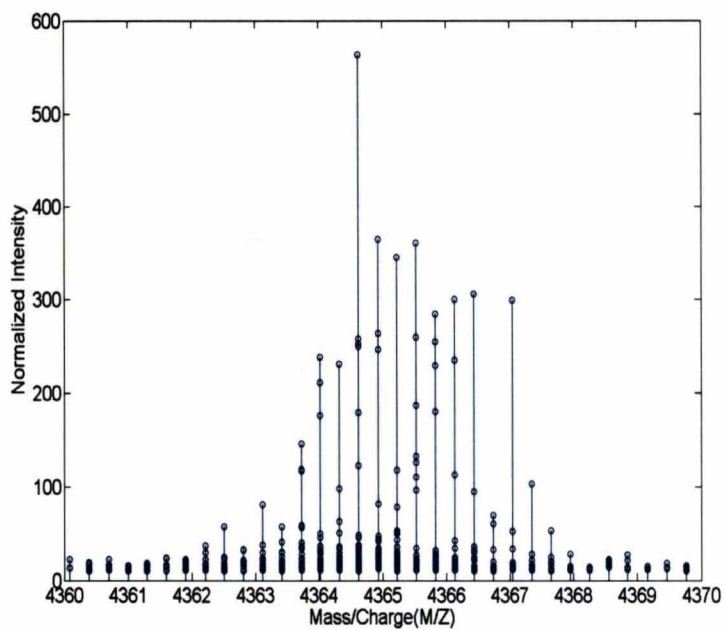
b)

Fig. 3.3[1] a) Projection of spectra onto a single axis and b) close-up of Fig. 3.3a.

---

[1]COURTESY: M. Vamsi et al. [14]
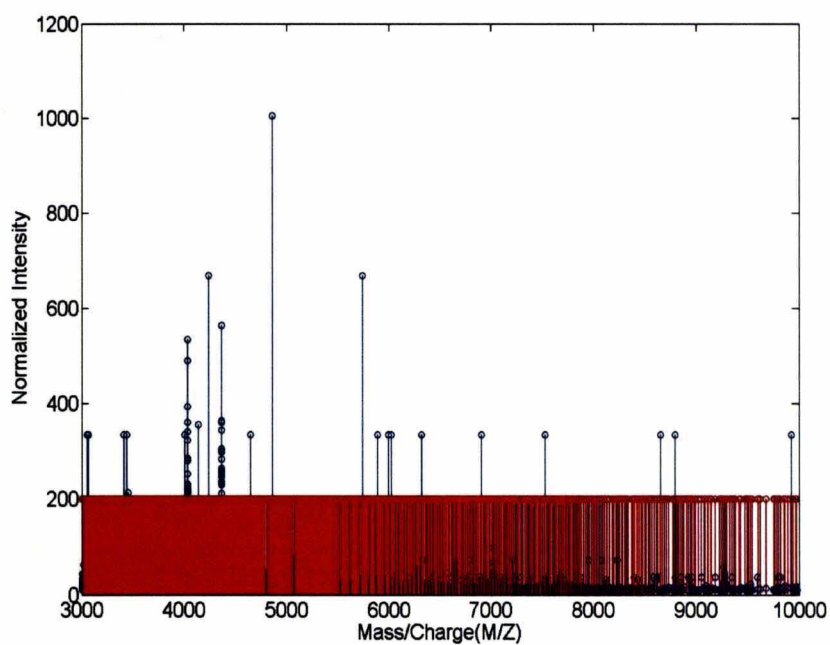
a)
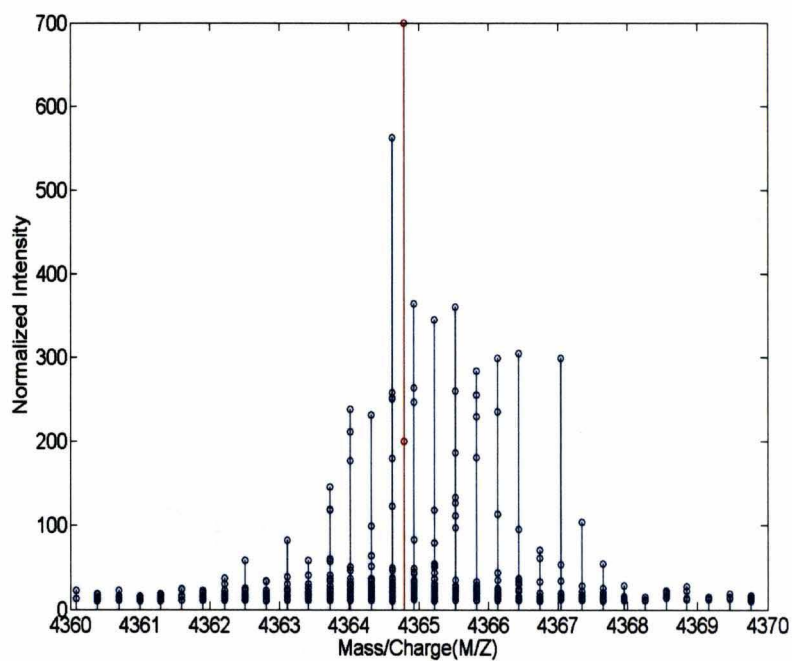
b)

Fig. 3.4[1] a) Result of clustering and **b)** close-up of Fig. 3.4a.
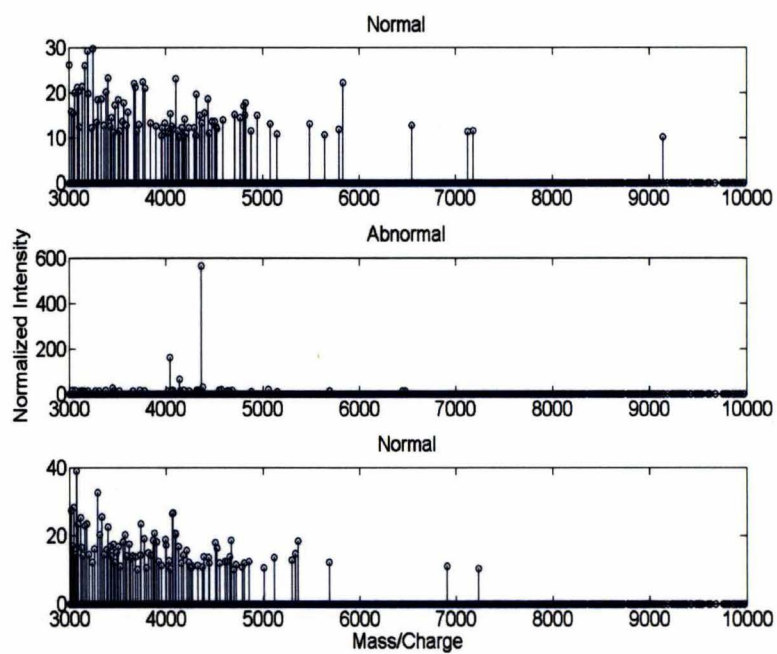
a)
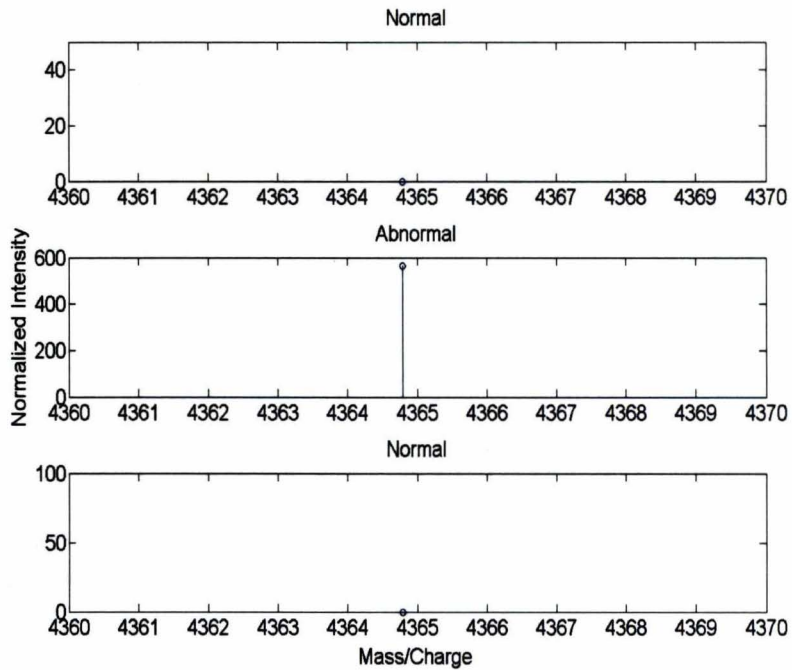
b)

Fig. 3.5[1] a) Result of the back projection and b) close-up of Fig. 3.5a.

3.2 Feature Selection Algorithm

The pre-processing steps remove the unnecessary peaks and all other unwanted noises. Now, we need to identify the candidate biomarkers or the optimal m/z values that differentiate the normal spots from cancer spots. From all the preprocessed mass spectra, hundreds of peaks were detected. We cannot utilize all of those peaks for classification as they could lead to problems like curse of dimensionality, slow convergence and large validation errors. To elude these problems, a feature selection algorithm was utilized to eliminate redundant peaks. This feature selection algorithm chose the best feature or the best combination of features that correlates to cancer. This algorithm consists of three important components: a piecewise linear classifier, an output reset algorithm and a floating search algorithm.

3.2.1 Piecewise Linear Classifier

The piece wise linear classifier (PLC) is a neural classifier [20], which is usually designed by minimizing the standard training error given by

$$E = \sum_{i=1}^{N_{class}} E(i) = \frac{1}{N_v} \sum_{i=1}^{N_{class}} \sum_{p=1}^{N_v} [t_p(i) - y_p(i)]^2 \tag{1}$$

where N_{class} denotes the number of classes, and $E(i)$ is the mean squared error for the ith output. Here $t_p(i)$ is the ith desired output for the pth pattern, and $y_p(i)$ is the ith observed output for the pth input pattern. N_v is the total number of data patterns. In the PLC, $y_p(i)$ is the output from the piecewise linear network.

$$y_p(i) = \sum_{j=1}^{N+1} w^{(q)}(i, j) x_p^{(q)}(j) \tag{2}$$

where N denotes the number of features; $w^{(q)}(i, j)$ is the model weight to the ith output from the jth feature in the qth cluster; $x_p^{(q)}(j)$ denotes the jth feature in the qth cluster; and $x_p^{(q)}(N+1)$ denotes the bias term, which is equal to one. The PLC works by approximating the general Bayes discriminant, by dividing the available data into a set of clusters and by solving a set of linear equations. We assume that $t_p(i_c) = 1$ and $t_p(i_d) = -1$, where i_c denotes the correct class number and i_d any incorrect class number for the current data pattern if $i_c = \mathbf{arg\ max}_i\ y_p(i)$. We say that PLC is correctly classified, otherwise an error is counted.

3.2.2 The output reset algorithm

The mean square error function given in (1) is not flexible, but confined. There are cases where the classification error remains unchanged with an increase or decrease in the error function. One such case is that if each individual output vector is added by a constant bias, the error function is either increased or decreased but keeping the classification error unaltered. Another case where there is no change in the classification error but having the error function increased or decreased is when an output has exact sign but the magnitude is larger than 1. In order to resolve these drawbacks and make the error function flexible, an output reset (OR) algorithm had been developed [21], in which the unwanted bias in the output vector y_p compensated by adding a constant a_p to each of the desired output vector. For the second case discussed above, where the output has the correct sign with a magnitude larger than 1, a function d_p is added to each of the desired vector. The revised error function becomes

$$E' = \frac{1}{N_v} \sum_{p=1}^{N_v} \sum_{i=1}^{N_{class}} [t'_p(i) - y_p(i)]^2 \qquad (3)$$

where $t_p{}'(i)=t_p(i)+a_p+d_p(i)$. All the training patterns are forced to become support vectors by the mean square error (MSE) type training, which can be alleviated by using OR. As our mass spectra consists of more normal spectra compared to cancer spectra, which are highly imbalanced, the OR algorithm proves to be extremely useful for the peak selection. The OR algorithm utilizes only some portions of the data, as the spectra far from the decision boundary will not be taken into consideration.

3.2.3 Floating search algorithm

The piecewise linear orthonormal least square (PLOLS) procedure is utilized to design the floating search algorithm [22]. The modified Schmidt procedure is utilized in PLOLS, making each feature in the cluster orthonormal. When the data is passed once, all the information that is required to search a good combination is stored in the auto- and cross- correlation matrices. This feature selection algorithm is very efficient and only one data pass is required. The modified desired output may be represented in a matrix form as

$$t' = x^{(q)}w^{(q)} + \Xi^{(q)} \tag{4}$$

where each row in matrix $\mathbf{x}^q$ represents one feature vector that is assigned to the qth cluster, $w^{(q)}$ denotes weight matrix in the qth cluster and $\Xi^{(q)}$ are residual errors in the qth cluster. The modified Schmidt procedure is applied to each cluster, yielding the piecewise linear orthogonal (PLO) system as,

$$t' = \Theta^{(q)}A^{(q)}w^{(q)} + \Xi^{(q)} = \Theta^{(q)}w_o^{(q)} + \Xi^{(q)} \tag{5}$$

3.2.4 Algorithm description

The feature selection algorithm which is proposed in the above selects N_s features from the available N features. This process can be described in the following steps.

1. PLC will use an appropriate number of clusters, N_c which is determined by trial and error method.

2. For each cluster solve a set of linear equations and design an N_c cluster PLC.

3. Using the OR algorithm, change the desired output.

4. By using the floating search algorithm, a good combination of features are selected.

3.3 Classification

Once the feature selection algorithm selected a compact set of a good combination of features, a multi layer perceptron (MLP) is used to classify the spectra as normal or cancer [23]. Unlike other classical objective functions, the classifier uses a new objective function with more free parameters to solve multiple sets of linear equations. The classifier uses an iterative minimization technique. An enhanced feed forward network reduces the error function with respect to the hidden weights. In this thesis, an OR integrated MLP is used, in which all the weights are subjected to optimization for further reducing the training error. In combination of OR enhanced MLP, an algorithm called output weight optimization-hidden weight optimization (OWO-HWO) [24] is also used. In OWO-HWO, the hidden unit weights and the output weights are modified alternately to reduce the training error.

CHAPTER 4

TEXTURE ANALYSIS

Texture defines the properties or characteristics of the surface of an object [25]. Texture analysis has been playing a key role in remote sensing and other surface inspection disciplines. It also extended its role in medical applications, typically by extracting the features automatically from the image that are utilized for further classification tasks. Sutton et al. [26] described the role of texture analysis in pulmonary disease classification. They have discussed three features of texture to discriminate normal lungs to that of diseased. The diagnosis of leukemic cancer with the help of texture properties is carried out by Harms et al. [27]. The ultra sound images of the heart were analyzed through the fractal texture features by Lundervold et al. [28].

Texture analysis can be performed through various methods, either statistical or syntactical. The former methods utilize the directionality and measure of coarseness, whereas the latter considers the shape and distribution of the cells [29]. David et al. [30] described a pyramid based texture analysis method that is based on a model of human texture perception. Many methods based on wavelets were proposed. A tree structure wavelet-transform based texture analysis is proposed by Tianhorng et al. [31]. Another method that uses the discrete wavelet transform to characterize the properties of texture is proposed by Unser et al. [32]. He used a discrete wavelet frame (DWF) i.e., an over-complete decomposition of the wavelets, in which no sub-sampling is done at the filter banks output. Haralick et al. [33] proposed the gray level co-occurrence matrix (GLCM) method, which is a statistical method of second order. Given an image having grayscale level in the range [0, N_g], where N_g is distinct number of gray levels,

matrix of size $(N_g+1)*(N_g+1)$ can be generated whose rows and columns represent the possible image value.

In this thesis, the work performed by Shao-Hui Chuang et al. [34] is utilized. The overview of the study can be seen in the system diagram shown in Fig. 4.1. The features used in this study are the gray level run-length matrix (GL-RLM) features. The study can be classified by main processes like feature extraction, feature selection and classification. A total of 13 features including 11 from the GL-RLM, mean and standard deviation are used. The GL-RLM features can be seen in the Table 4.1. The final feature matrix is obtained by summing up all the four directional feature matrices. The feature selection is carried out by a piecewise linear network model [22]. Finally, a classification algorithm that is the same as the one used in the MALDI-MS method is used.

Table. 4.1. The features used for texture analysis [34]

Feature	Formula
Short Run Emphasis	$SRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{j^2}$
Long Run Emphasis	$LRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j)*j^2$
Low Gray-Level Run Emphasis	$LGRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2}$
Hign Gray-Level Run Emphasis	$HGRE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j)*i^2$
Short Run Low Gray-Level Emphasis	$SRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)}{i^2 * j^2}$
Short Run High Gray-Level Emphasis	$SRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)*i^2}{j^2}$
Long Run Low Gray-Level Emphasis	$LRLGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{p(i,j)*j^2}{i^2}$
Long Run High Gray-Level Emphasis	$LRHGE = \frac{1}{n_r} \sum_{i=1}^{M} \sum_{j=1}^{N} p(i,j)*i^2 * j^2$
Gray-Level Non-uniformity	$GLNU = \frac{1}{n_r} \sum_{i=1}^{M} \left(\sum_{j=1}^{N} p(i,j) \right)^2$
Run Length Non-uniformity	$RLNU = \frac{1}{n_r} \sum_{j=1}^{N} \left(\sum_{i=1}^{M} p(i,j) \right)^2$
Run Percentage	$RPC = \frac{n_r}{M*N}$
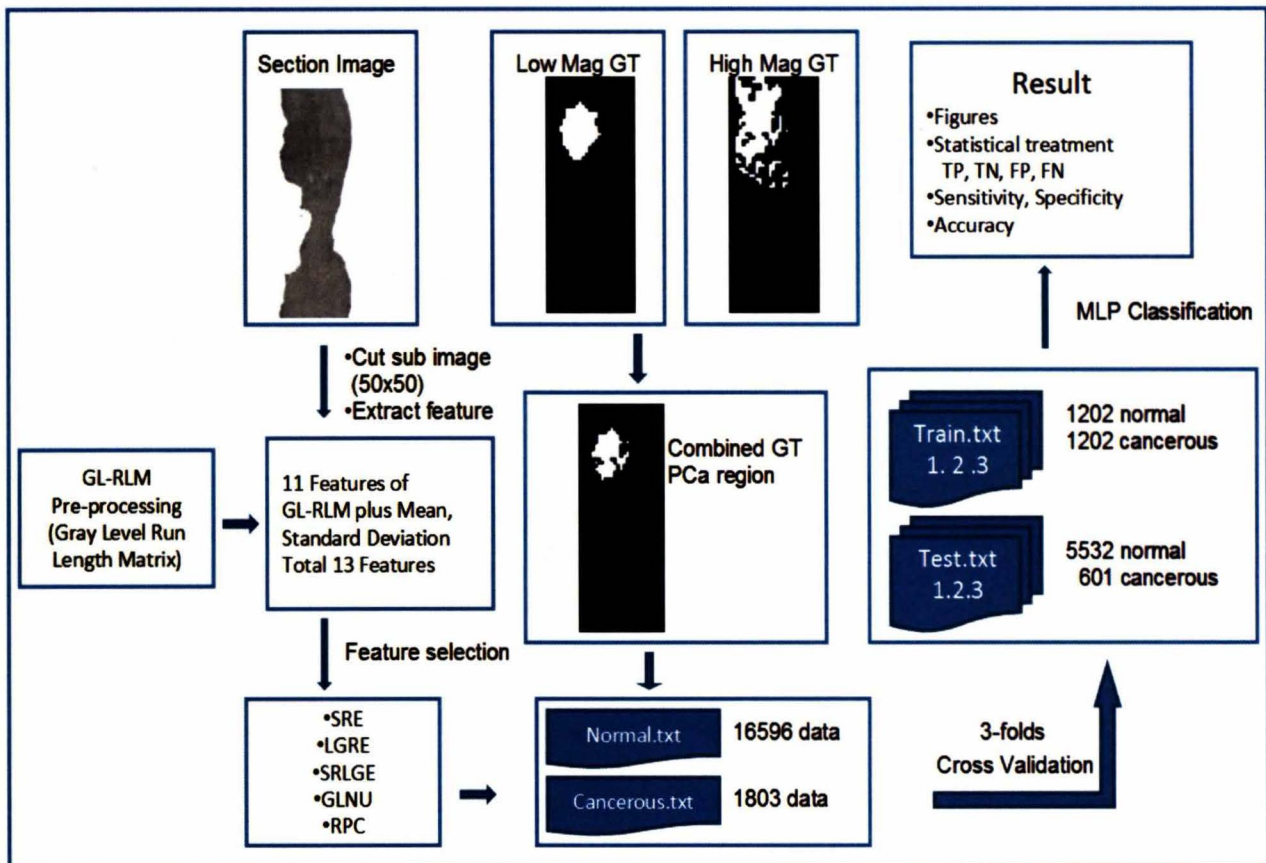
Fig. 4.1[2] System overview of texture analysis

CHAPTER 5

PERFORMANCE METRICS

This chapter discusses the metrics that are used to evaluate the performance of our study and provide some numbers to define the efficiency.

5.1 K-fold Cross Validation

The k-fold cross validation method is said to be an improved hold out method. In this method, the original data is randomly divided into k parts. Of all the parts, one is used for validation (testing) and all remaining parts are used for training. The training and testing is repeated k times so that every part is used as validation once. A single estimate can be produced by either combining or averaging the k validation results. An example for the k-fold cross validation can be seen in Fig. 5.1. Here k is 3.

5.2 Sensitivity and Specificity

For a binary classification problem, the performance can be measured through sensitivity and specificity. The sensitivity measures the correctly classified positives, and the specificity calculates the correctly classified negatives.

$$sensitivity = \frac{numberofTruePositives}{numberofTruePositives + numberofFalseNegatives}$$

$$specificity = \frac{numberofTrueNegatives}{numberofTrueNegatives + numberofFalsePostives}$$

5.3 Confusion Matrix Table

The sensitivity and specificity calculated for an experiment can be shown through a table called confusion matrix table. It also shows the accuracy of the experiment conducted. A confusion matrix table can be seen in Fig. 5.2.

5.4 Hirox KH-1300

The machine used for our study is the Hirox KH-1300. It has a high resolution microscope and a PC connected to it. The microscope is set to a magnification of 420x. The image captured under the microscope can be transferred to the PC at high speed. The E-tiling software used to reconstruct the images is built-in software in the Hi-Scope KH-1300.
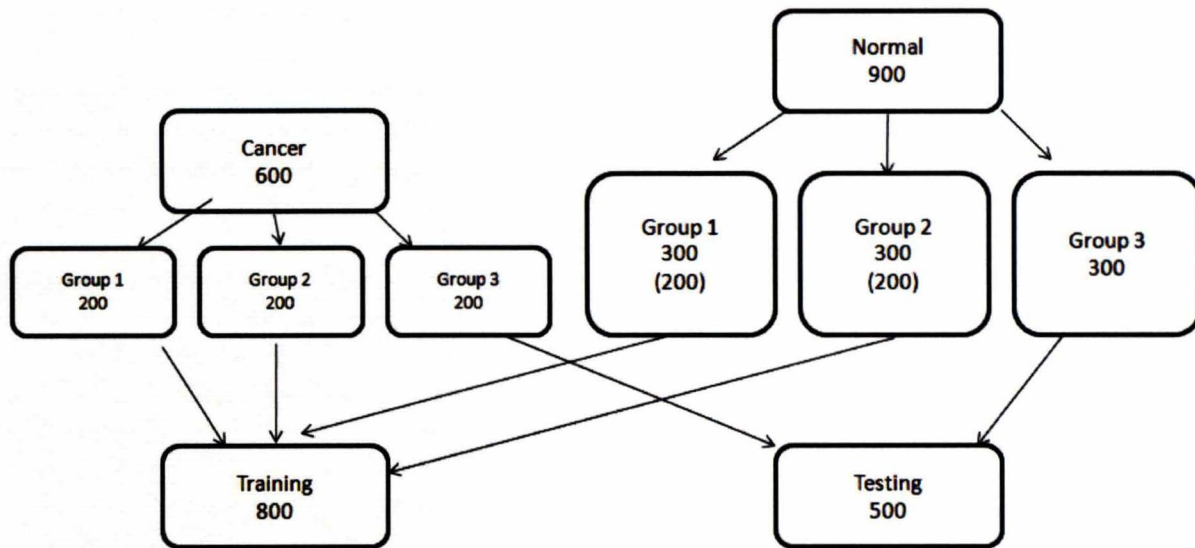
Fig. 5.1 An example for 3-fold cross validation

		Actual Cancer Condition	
		TRUE (Cancer)	**FALSE (Normal)**
Predicted Result	**Positive (Cancer)**	True Positive	False Positive
	Negative (Normal)	False Negative	True Negative
		Sensitivity $TP/(TP+FN)$	**Specificity** $TN/(TN+FP)$
		Accuracy $(TP+TN)/(TP+TN+FP+FN)$	

Fig. 5.2 Confusion matrix table

CHAPTER 6

EXPERIMENTS AND RESULTS

This chapter provides the information on the experiments performed in our study. Results are also given to check the efficiency of the proposed model.

6.1 Experiments

In this thesis, three experiments are conducted. The experiments were performed to analyze the efficiency of the MALDI mass spectrometry in predicting the true tumor region marked by the pathologist as highly confident cancer regions.

Experiment A: Cancer MS spectra for training were taken from the intersection of the cancer areas in AI1 and AI2 , and normal MS spectra for training were taken from a normal area that are not close to the cancer spots, yielding 19 cancer and 19 normal spectra in the training data. Feature selection and classification are then performed on those preprocessed training spectra, yielding a classifier model. The trained model was subsequently applied to AI2 to obtain testing classification accuracy.

Experiment B: Cancer MS spectra for training were taken from the cancer areas in AI1, and normal MS spectra for training were obtained in a similar way as that in experiment A. The training data contained 27 cancer and 27 normal spectra. After feature selection and classification, a trained model was applied to AI2 to get testing performances. This is the most important experiment, because the overall goal of this study is to predict a more accurate cancer boundary based on a low resolution ground truth and to aid in the biomarker identification task.

Experiment C: A 3-fold cross validation (CV) is performed using the pathologist annotated cancer and normal regions in AI2 as ground truth. There were 51 cancer and 923 normal spectra in the data set. In the 3-fold CV procedure, data was partitioned into 3 parts; 2 parts were used for training, and the remaining part was used for testing. This procedure was repeated three times such that each part was used for the testing once. The tested parts were then pooled to compute sensitivity and specificity for the classification. Similar experiments had been performed in texture analysis in [34].

6.2 Results

For MALDI, sensitivities and specificities were calculated for each of the experiments and are shown in Table 6.1. The feature selection algorithm selected 5, 5 and 3 peaks for classification in experiments A, B and C, respectively. In experiment A, a sensitivity of 50.92% and a specificity of 98.65% is achieved. In experiment B, a sensitivity of 54.90% and a specificity of 98.48% is obtained. In experiment C, a sensitivity of 62.75% and a specificity of 98.37% were achieved. These prediction results were mapped back to the MALDI data and are shown in Fig. 6.1a, 6.1b and 6.1 c. For texture analysis, the results were better than the MALDI MS. In experiment A, a sensitivity of 83.92% and a specificity of 81.25% are achieved. In experiment B, a sensitivity of 87.45% and a specificity of 75.00% were obtained. In the last experiment C, a sensitivity of 95.29% and a specificity of 72.03% were achieved. These can be clearly seen in the table 6.2. These predictions were mapped back to the artificial histopathological tissue image and are shown in Fig. 6.1d, 6.1e and 6.1f.

The above results show the poor prediction of tumor region through MALDI mass spectrometry, though it obtained a very high specificity. This implies an inadequate prediction of

the true cancer region (true positives). The texture analysis results are viceversa, having good sensitivities and relatively poor specificities.

Table. 6.1. Results of MALDI MS

Experiments	Sensitivity	Specificity	Accuracy
Exp A	51.02%	99.67%	97.12%
Exp B	50.98%	100.00%	97.43%
Exp C	58.82%	96.82%	96.82%

Table. 6.2. Results of texture analysis

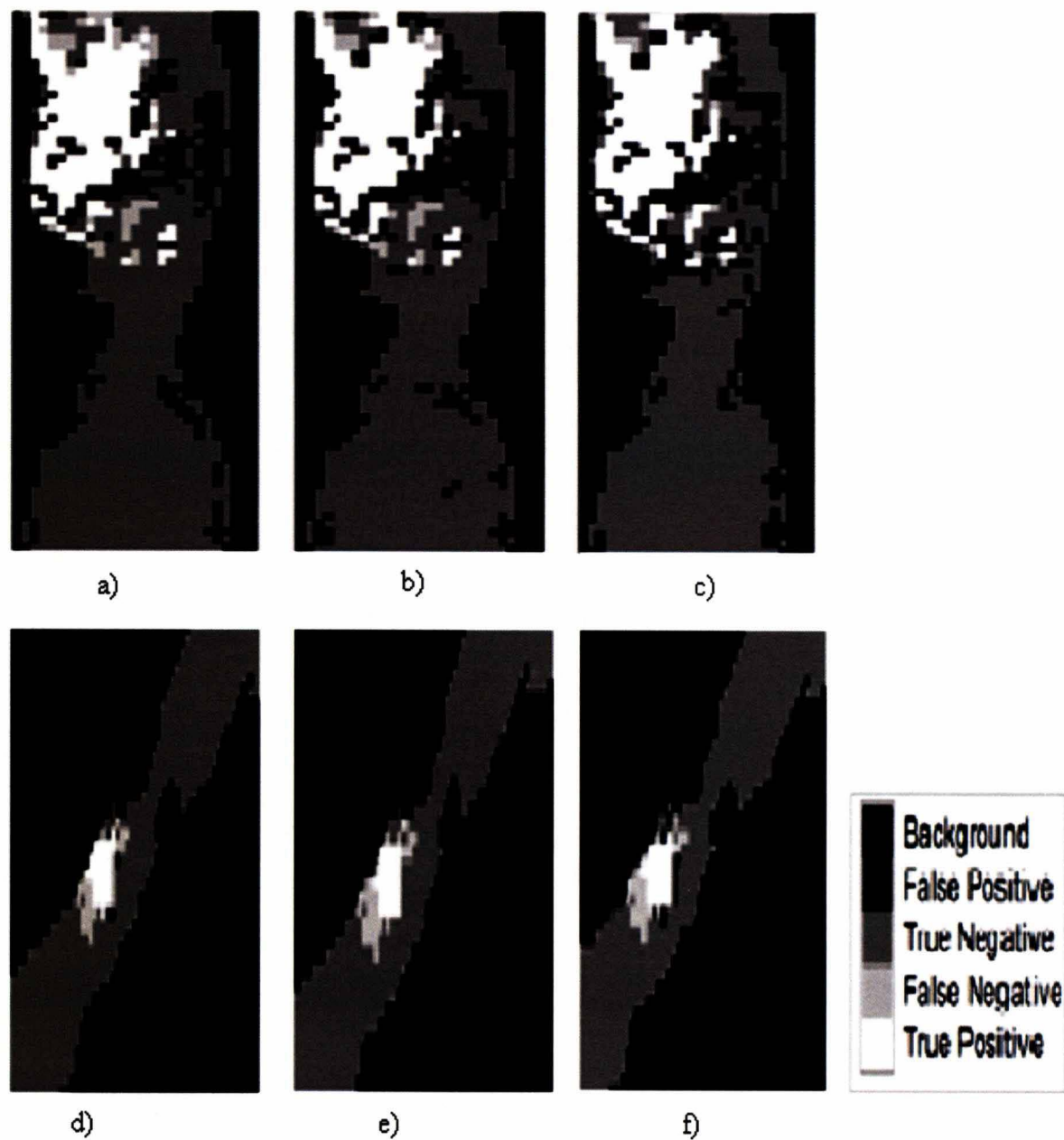| Experiments	Sensitivity	Specificity	Accuracy
Exp A	83.92%	81.25%	81.82%
Exp B	87.45%	75.00%	77.65%
Exp C	95.29%	72.03%	76.98%

Fig. 6.1 Texture Analysis: a) Result of experiment A, b) Result of experiment B and c) Result of experiment C. MALDI: d) Result of experiment A, e) Result of experiment B and f) Result of experiment C.

CHAPTER 7

FUSION OF RESULTS

This chapter discusses the third step of our proposed work, which combines the results from MALDI MS processing and texture analysis methods.

7.1 Fusing the Results

The MALDI MS processing results are low in sensitivities and high in specificities, while the texture analysis results are good in sensitivities and relatively poor in the specificities. In an efficient model for diagnosis, optimal sensitivities and specificities are needed. To design an efficient model that can predict a true tumor boundary, results from both the texture analysis and the MALDI MS processing are fused. The specimen tissues used in the two methods are adjacent slices, which are almost identical and landmark registrations had been performed to ensure the similarity. The 200*200 block used in the texture analysis corresponds to a single spot (spectrum) in the MALDI. The classification probabilities from the two techniques were averaged to get a new probability. A classifier relies on the classification probabilities to discriminate normal from cancer. Typically, 0.5 is put as a threshold and if the probability is greater (lesser) than the threshold, it gives a 1(2), where 1 and 2 refer to normal and cancer (or vice versa). For example, if the classification of the texture analysis yields a probability of 0.3 (normal) and yields a probability of 0.9 (cancer) from the MALDI MS, the average of these two probabilities gives a 0.6, which results in a classification of cancer. Similarly, if one method gives a probability of 0.5 (cancer) and another method yields a probability of 0.4 (normal), the fusion of these two probabilities makes the classification to be a normal one. Thus, the averaging

of the classification probabilities from both the techniques yielded surprisingly good results, which are better in both the sensitivities and specificities. The crucial experiment B yielded a sensitivity of 80.39% and a specificity of 93.09%. Results are shown in Table 7.1 and in Fig. 7.1 and Fig. 8.2.

Table 7.1. Results from fusion

Experiments	Sensitivity	Specificity	Accuracy
Exp A	84.31%	87.83%	87.32%
Exp B	80.39%	93.09%	91.26%
Exp C	88.29%	85.13%	85.64%
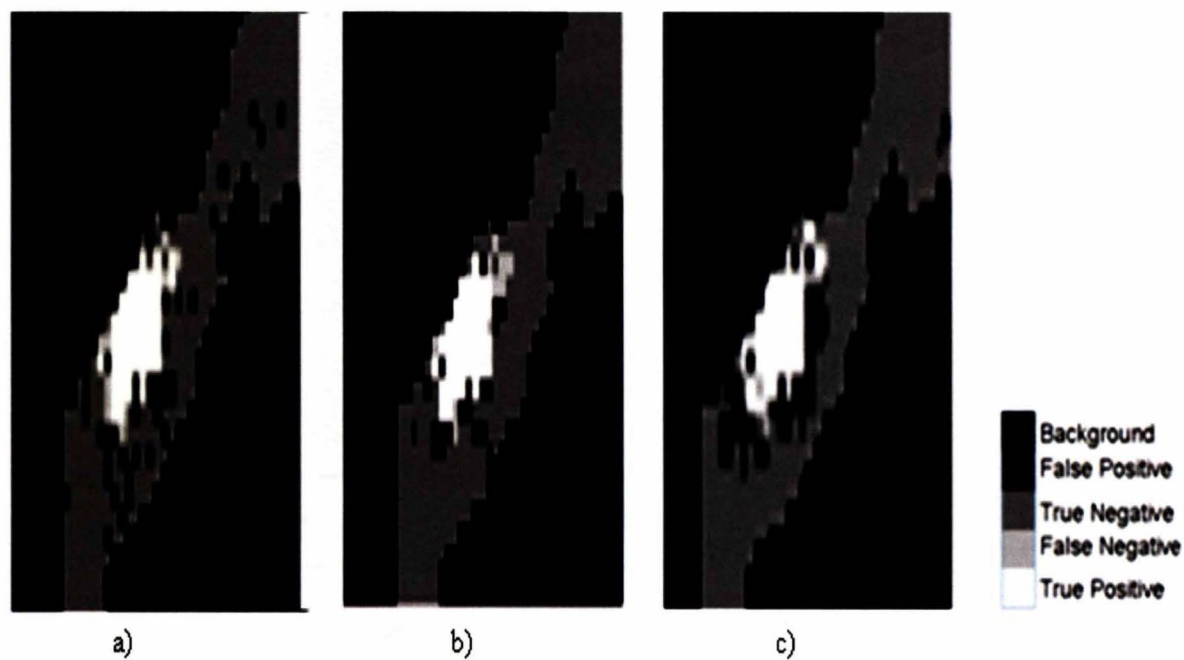
Fig. 7.1 a) Result of experiment A after the fusion, b) Result of experiment B after the fusion and c) Result of experiment C after the fusion.
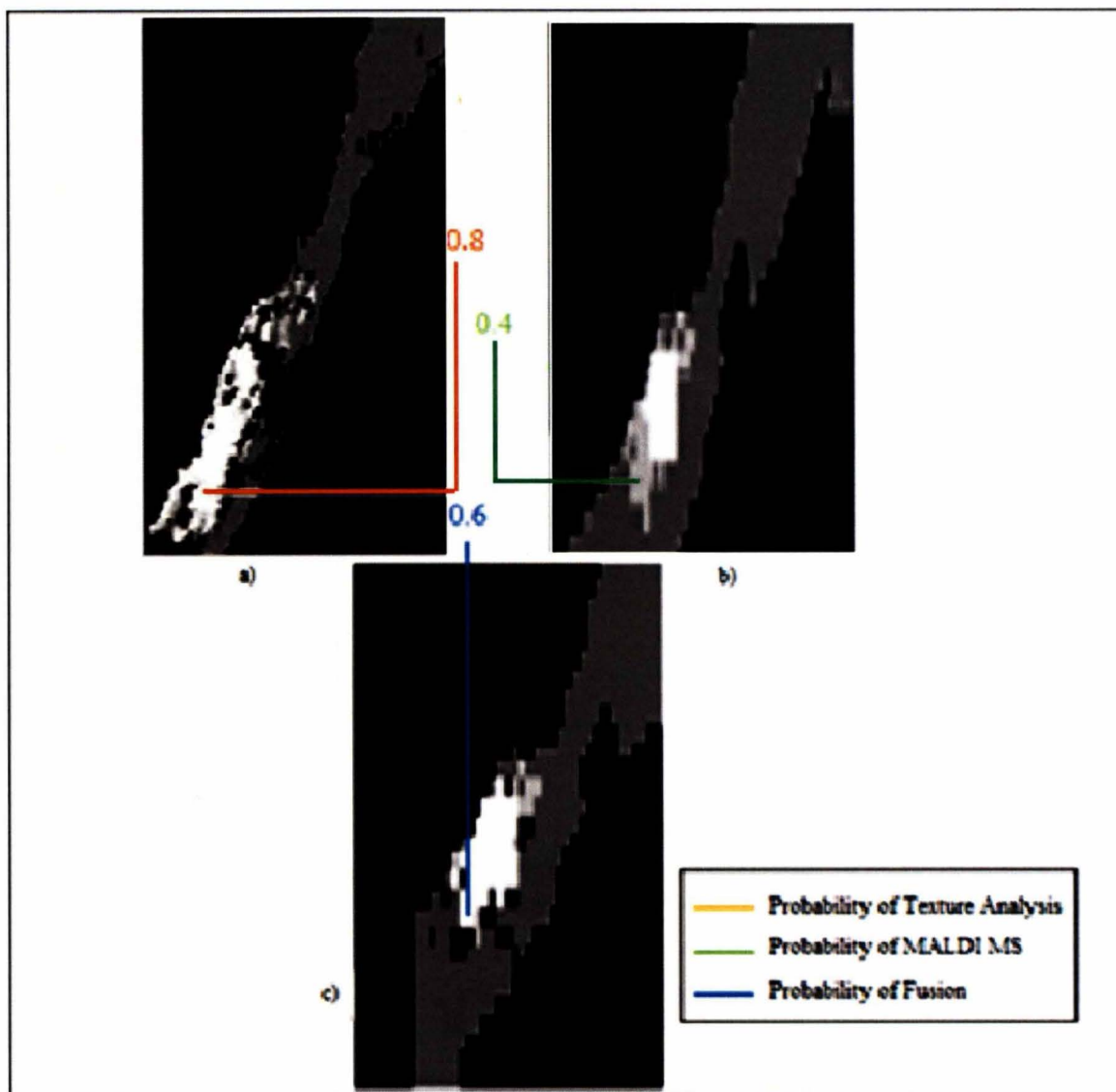
Fig. 7.2 a) Result of texture analysis showing a given point as cancer, b) Result of MALDI showing the same given point as normal and c) Result of fusion showing the point as cancer (close-up).

7.2 Discussion

The proposed methods aim at confronting the MALDI mass spectrometry in predicting the true tumor boundary. The high confidence area marked by the pathologist was taken as the reference (true cancer region). The role of molecular biomarkers in the estimation of the tumor

boundary appears to be inadequate from the obtained sensitivities and specificities. Poor sensitivity shows the inability to find the true cancer region. On the contrary, a very high specificity is achieved, which shows there is no normal region that has been diagnosed as cancer (no false positives). On the other hand, the texture analysis results that were conducted parallel on optical images from an adjacent slice proved to be a better estimate for the sensitivity. To enhance the efficiency of the molecular biomarkers, the imaging biomarkers are combined with these molecular biomarkers. The fusion process produced much better results that are very good in both sensitivity and specificity.

CHAPTER 8

CONCLUSIONS

In this thesis, PCa region prediction models were developed by fusing results from a texture analysis technique and predictions from MALDI spectra processing. The designed models are able to predict a high resolution PCa region based on a low-resolution PCa area defined on the adjacent biopsy slice. This prediction methodology can benefit prostate cancer biomarker identification using MADLI imaging techniques by providing a more realistic indication of where PCa regions are located including those regions too difficult to identify without painstaking and time intensive histology analysis and impossible to identify on the non-histologically processed adjacent slice subjected to MALDI proteomic processing. Such regions may also contain useful biomarker information, especially at the PCa borders.

REFERENCES

[1]   Prostate cancer foundation (PCF),

        http://www.prostatecancerfoundation.org/site/c.itIWK2OSG/b.4983495/k.5C76/About_Pros

        tate_Cancer.htm, Date accessed February 2010.

[2]   American Cancer Society Inc.,

        http://www.cancer.org/docroot/CRI/content/CRI_2_2_1X_How_many_men_get_prostate_c

        ancer_36.asp?sitearea, Date accessed April 2010.

[3]   P. Paola, B. Freddie, and P. D. Maxwell, "Estimates of the world-wide prevalence of cancer

        for 25 sites in the adult population," *International Journal of Cancer*, vol. 97, no. 1, pp. 72-

        81, 1997.

[4]   National Cancer Institute (NCI),

        http://www.cancer.gov/cancertopics/factsheet/Detection/PSA, Date accessed February 2010.

[5]   W. J. Catalona, D. S. Smith, T. L. Ratliff, K. M. Dodds, D. E. Coplen, J. J. Yuan, J. A.

        Petros, and G. L. Andiole, "Measurement of prostate-specific antigen in serum as a

        screening test for prostate cancer," *New England Journal of Medicine*, vol. 324, pp. 1156–

        1161, 1991.

[6]   M. Hardt, et al.,"Toward defining the human parotid gland salivary proteome and

        peptidome: identification and characterization using 2D SDS-PAGE, ultrafiltration, HPLC,

        and mass spectrometry," *Biochemistry*, vol. 44, pp. 2885–2899, 2005.

[7]   X. Deng, H. Geng and H. H. Ali, "Cross-platform analysis of cancer biomarkers: A bayesian network approach to incorporating mass spectrometry and microarray data," *Cancer Informatics*, vol. 3, pp. 183–202, 2007.

[8]   C. Laronga, and R. R. Drake, "Proteomic approach to breast cancer," *Cancer Control*, vol. 14, no. 4, pp. 360-368, 2007.

[9]   Biomarkers definitions working group, "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework," *Clinical Pharmacology and Therapeutics*, vol. 69, pp. 89–95, 2001.

[10]   J. Hanley, B. McNeil, "The meaning and use of the area under a receiver operating characteristic curve," *Diagnostic Radiology*, vol. 143, no. 1, pp. 29-36, 1982.

[11]   H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226-1238, 2005.

[12]   L. Li, D. M. Umbach, P. Terry and J. A. Taylor, "Application of the GA/KNN method to SELDI proteomics data," *Bioinformatics*, vol. 20, no. 10, pp. 1638-1640, 2004.

[13]   M. Vamsi, J. Li and F. McKenzie, "Biomarker identification by using mass spectrum data," *Capstone conference*, presentation, VMASC 2008.

[14]   M. Vamsi, "Biomarker identification for prostate cancer using an efficient feature selection algorithm," Thesis, Old Dominion University, December 2008.

[15]   MALDI-MSI interest group,

http://www.maldi-msi.org/index.php?option=com_content&task=view&id=14&Itemid=39,

Date accessed December 2009.

[16]   The Mathworks Inc., http://www.mathworks.com/products/product_listing/index.html, Date

accessed January 2009.

[17]   W. M. Old, et al., "Comparison of label-free methods for quantifying human proteins by

shotgun proteomics," *Molecular Cellular Proteomics*, vol. 4, no. 10, pp. 1487-502, 2005.

[18]   K. W. Jordan, W. He, E. F. Halpern, C. Wu and L. L. Cheng, "Evaluation of tissue

metabolites with high resolution magic angle spinning MR spectroscopy human prostate

samples after three-year Storage at -80 °C," *Biomarker Insights*, vol. 2, pp. 147–154, 2007.

[19]   B. Adam, et al.,"Serum protein fingerprinting coupled with a pattern-matching algorithm

distinguishes prostate cancer from benign prostate hyperplasia and healthy men," *Cancer

Research*, vol. 62, pp. 3609-3614, 2002.

[20]   A. A. Abdurrab, M. T. Manry, J. Li, S. S. Malalur, and   R.G. Gore, "A piecewise linear

network classifier," in *Proceedings of International Joint Conference on Neural Networks*,

Orlando, Florida, pp. 1750-1755, 2007.

[21]   J. Li, M. T. Manry, L. Liu, C. Yu, and John Wei, "Iterative improvement of neural

classifiers," in *Proceedings of the Seventeenth International Conference of the Florida AI

Research Society*, pp. 700-705, 2004.

[22]   J. Li, J. Yao, R. M. Summers, N. Petrick, M. T. Manry, and A. K. Hara, "An efficient feature selection algorithm for computer-aided polyp detection," *International Journal on Artificial Intelligence Tools*, vol. 15, no. 6, pp. 893–915, 2006.

[23]   R. G. Gore, J. Li, M. T. Manry, L. M. Liu, and C. Yu, "Iterative design of neural network classifiers through regression," *Special Issue of International Journal on Artificial Intelligence Tools*, vol. 14, no. 1-2, pp. 281–302, 2005.

[24]   C. Yu, and M. T. Manry, "A modified hidden weight optimization algorithm for feed-forward neural networks," *Thirty-Sixth Asilomar Conference on Signals, Systems & Computers*, pp. 1034-1038, 2002.

[25]   X. Tang, "Texture information in run-length matrices," *IEEE Transaction on Image Processing*, vol. 7, no. 11, pp. 1602-1609, 1998.

[26]   R. Sutton and E. L. Hall, "Texture measures for automatic classification of pulmonary disease," *IEEE Transactions on Computers*, C-21, pp. 667-676, 1972.

[27]   H. Harms, U. Gunzer, and H. M. Aus, "Combined local color and texture analysis of stained cells," *Computer Vision, Graphics and Image Processing*, 33, pp. 364-376, 1986.

[28]   G. H. Landeweerd and E. S. Gelsema, "The use of nuclear texture parameters in the automatic analysis of leukocytes," *Pattern Recognition*, vol. 10, pp. 57-61, 1978.

[29]   F. Argenti, L. Alparone and G. Benelli, "Fast algorithms for texture analysis using co-occurence matrices," in *Proceedings of IEEE*, vol. 137, pt. F, No.6, pp. 443-448, 1990.

[30]   D. Heeger, J. Bergen, "Pyramid based texture analysis/synthesis," *ACM Proceedings*, 0-89791-701, pp. 229-238, 1995.

[31]   T. Chang and C. C. Jay Kuo, "Texture analysis and classification with tree-structured wavelet transform," in *Proceedings of IEEE Conference on Image Processing*, vol. 2, no. 4, pp. 429-441, 1993.

[32]   M. Unser, "Texture classification and segmentation using wavelet frames," in *proceedings of IEEE Conference on Image Processing*, vol 4, pp.1549-1560, 1995.

[33]   R. Shanmugam, "Textural features for image classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 3, pp. 610-621, 1973.

[34]   S. Chuang, et al., "Adjacent slice prostate cancer prediction to inform MALDI imaging biomarker analysis," in *Pproceedings of SPIE*, vol. 7624, pp. 433-441, 2010.