Fall 2008

Biomarker Identification for Prostate Cancer Using an Efficient Feature Selection Algorithm

Vamsi Krishnam Raju Mantena
Old Dominion University

BIOMARKER IDENTIFICATION FOR PROSTATE CANCER USING AN EFFICIENT FEATURE SELECTION ALGORITHM

by

Vamsi Krishnam Raju Mantena
B.E., April 2006, Andhra 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
December 2008

Approved by:

Jiang Li (Director)

Yuzhong Shen (Member)

Frederic McKenzie (Member)

ABSTRACT

BIOMARKER IDENTIFICATION FOR PROSTATE CANCER USING AN EFFICIENT FEATURE SELECTION ALGORITHM

Vamsi Krishnam Raju Mantena
Old Dominion University, December 2008
Director: Dr. Jiang Li

In recent years, there has been an increased interest in using protein mass spectrometry to identify biomarkers that discriminate diseased from healthy individuals. A biomarker is a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathological processes, or pharmacological responses to a therapeutic intervention. Identifying biomarkers will be an important step towards disease characterization and patient management. One challenge of biomarker identification is how to handle the high dimensional mass spectral data. In this thesis, we applied an efficient feature selection algorithm to mass spectrometry data obtained from prostate tissue samples to identify prostate specific cancer biomarkers. Experiments showed that the proposed method achieved high sensitivities and specificities and outperformed many other currently used feature selection algorithms.

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 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 Dr. John Semmes at East Virginia Medical School (EVMS) for providing the mass spectrum data 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

Chapter Page

Chapter Page

LIST OF FIGURES

Figure Page

Figure Page

LIST OF TABLES

Table Page

CHAPTER 1

INTRODUCTION

1.1 Background

Early detection of various cancers is very important because many cancers are treatable if they are diagnosed in the early stages. For example, prostate cancer diagnosed while still localized to the prostate can be cured by a number of local therapies. Prostate cancer is a disease in which cancer develops in the prostate, a gland in the male reproductive system. It occurs when cells of the prostate mutate and begin to multiply out of control. Rates of prostate cancer vary widely across the world. Although the rates vary widely between countries, it is least common in South and East Asia, more common in Europe, and most common in the United States [1]. Prostate cancer develops most frequently in men over fifty in the United States. It is responsible for more male deaths than any other cancer except lung cancer. In the UK it is also the second most common cause of cancer death after lung cancer. Around 35,000 men in the UK are diagnosed per year and around 10,000 die of it.

Prostate-specific antigen (PSA) in serum is currently the most popular approach for its early detection. A high PSA concentration implies a possible high risk of prostate cancer. However, because PSA is prostate specific rather than prostate cancer specific, increased concentrations of PSA are also found in benign prostatic hyperplasia [2], acute and chronic prostatitis, and prostatic intraepithelial neoplasia [3]. On the other hand, previous studies reported that 15% of men will have prostate cancer even when their PSA concentrations are low [4-5]. Several approaches have been undertaken to improve the

---

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

PSA test such as measuring PSA velocity, PSA density, and assessing ratios between free, complexed, and total PSA serum values with various degrees of success. Combinations of markers such as free PSA, IGF-I, and IGF-binding protein 3 have resulted in improved diagnostic discrimination between benign prostatic hyperplasia (BPH) and prostate cancer [6]. Suspected prostate cancer is typically confirmed by taking a biopsy of the prostate and examining it under a microscope. Further tests, such as CT scans and bone scans, may be performed to determine whether prostate cancer has spread. It is becoming increasingly clear that because of the inherent molecular heterogeneity and multifocal nature of prostate cancer, additional improvement in early detection, diagnosis, and prognosis will likely require the measurement of a panel of biomarkers [6].

1.2 Biomarkers

Biomarkers, in the context of cancer diagnosis, usually refer to specific genes and their products, which are indicators of disease states and can be detected in clinical settings [7]. The discovered biomarker should possess key characteristics and qualities that will depend upon its intended use. A biomarker must be accurate, sensitive and specific [8]. It should be able to discriminate between diseased and controlled populations [9-10], quantified reliably and reproducibly.

1.3 Mass Spectrometry

In recent years, technological developments have spurred interests in using protein mass spectrometry to identify molecular markers for discriminating between

phenotypic groups [11]. The diagnostic categories often consist of tumor versus normal tissues, different types of malignancies, and subtypes of a specific cancer. There are often hundreds of peaks (protein biomarkers) detected from a set of mass spectra. Using all of the detected peaks for the classification can have side effects including the curse of dimensionality, convergence difficulties and large validation errors. In order to avoid these problems, feature selection [12] is normally utilized to generate a compact subset of peaks that leads to accurate models for the available data. Microarrays and mass spectrometry, a pair of complementary tools for studying genome activity and proteome activity respectively, have emerged to bring hopes for discovering biomarkers and building diagnosis models.

Mass spectrometers share at least three common features: an ionization source, a mass analyzer, and a detector. For analysis of proteins, particularly from clinical samples, the two most commonly used mass spectrometers involve surface-enhanced laser desorption/ionization (SELDI) and matrix-assisted laser desorption/ionization (MALDI) sources with a time-of-flight (TOF) mass analyzer [13]. The MALDI-TOF approach is more amenable to the higher throughput analysis of many clinical specimens, as the laser desorption process results in primarily single ion species, allowing for a profiling approach of multiple species.

A mass spectrum can be represented by 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 corresponding *m/z* value. Mass spectra are difficult to interpret; a considerable amount of time and effort should be spent in spectra preprocessing and peak feature selection in order to improve their quality and reduce

their dimensionality. The goal of the analysis is often to identify peaks related to specific outcomes, such as different malignancies or clinical responses such as cancer. Recent progress in mass spectrometry has shown the promising potential of biomarker discovery in the diagnosis of diseases especially in the early stages.

1.4 Early Detection Research Network (EDRN)

In 2000, the National Cancer Institute organized The Early Detection Research Network (EDRN) to bring together institutions to help accelerate biomarker research. A biomarker is defined as "A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathological processes, or pharmacological responses to a therapeutic intervention [14]." Though biomarker research progressed considerably in recent years, the practical impact of this research on screening, diagnosis and prognosis remains limited, partially due to the large number of biomarkers available. None gets approved by regulatory bodies, and none seems to be capable to point the way to specific therapies [15].

Recent projects initialized by EDRN for validating the previously identified biomarkers showed that few of the serum biomarkers do not reliably detect prostate cancer [16-17]. The projects consist of 3 stages that were targeted at evaluation of the previously published EDRN study for the detection of prostate cancer. In stage 1, a group of institutions reported that SELDI-TOF mass spectrometry instruments at separate sites could be accurately standardized over a 3-month period and could be used to accurately classify previously studied prostate cancer patient and control sera using known spectral features [18]. In stage 2, the aim was to develop the original algorithm and to test the

developed algorithm on a set of independent patients. In this stage, the validation failed and bias was identified in the patient samples [16]. The experiment was then redesigned and patient's sample bias was eliminated. It was found that the redesigned classifier again failed to separate patients with prostate cancer from biopsy-negative controls, nor did it separate patients with prostate cancer with Gleason scores < 7 from those with Gleason scores ≥7 [17]. Stage 3 is a prospective study that was not performed because the validation failed in stage 2. In stage 2, classification performance given by the constructed classifier were highly unbalanced for many cases, i.e., sensitivities were much higher than specificities or vice versa. The overall accuracy was close to what it would be if all samples were classified as prostate cancer.

There are four possible reasons for the failure. First, SELDI-TOF MS whole serum proteomic profiling might not be powerful enough to reliably detect prostate cancer as stated in [17]. Second, classifiers used in the experiment were a forward stepwise logistic regression and a boosted logistic regression. Those two particular classifiers may not be the best for this task. Different applications require particularly tailored algorithms to deal with specific challenges faced by the problem [19]. Third, only the area under the ROC curve method was utilized for biomarker identification (peak selection) before classifier construction, peaks were further selected in the classifier design. These peak selection algorithms may not be powerful enough to identify the best peak combination for the classification. Finally, the peak lists detected by different institutions may not be well calibrated. The aim of this thesis is to test whether one of the feature selection algorithms [20] developed by our team can improve prostate cancer detection accuracy.

1.5 Challenges faced by the Existing Feature Selection Algorithms for Biomarker Identification

Several feature selection algorithms have been applied to biomarker identification. Nadedge Dossat *et al.* (2007) utilized the Wilcoxons test to preselect a set of peaks detected from surface enhanced laser desorption/ionization time-of-flight (SELDI-TOF) spectra before applying various classifiers for cancer diagnosis [21]. James Lyons-Weilera *et al.* (2005) [14] reviewed more feature selection algorithms that have been used for biomarker identification, including Area under ROC curve, Fisher score, J5 test, simple separability criterion, *t*-test score and weighted separability criterion. After biomarker selection, they applied a de-correlation step to delete the selected but correlated peaks. The de-correlation step will improve subsequent classification with those redundant peaks being removed. However, those peak selection algorithms mainly perform the selection on each individual peak, and important interactions among peaks may be missed.

Yinsheng Qu *et al.* (2003) [22] proposed a wavelet based algorithm to reduce the number of peaks. After the wavelet transformation, less than 10 wavelet coefficients have been utilized in classification. Using wavelet transformation to reduce the dimensionality might be useful for classification but has less importance for biomarker identification. A wavelet coefficient is a linear combination of all available peaks, and the most discriminative wavelet coefficient is not a biomarker but a particular combination of all peaks. Principle component analysis falls into the same category. A genetic algorithm [23] is a good candidate for biomarker identification and is embedded in some mass spectrometry software (For example, CLINPROT). However, a genetic algorithm needs

many user defined parameters including the number of peaks you want to select, which is hard to determine beforehand.

1.6 Proposed Work

The proposed method starts by applying the data preprocessing techniques to the raw mass spectrum data. Data preprocessing techniques [24] such as baseline adjustment, smoothing, normalization, peak detection and clustering [25] improve the performance of mass spectrometric data analysis methods for biomarker discovery. The reason for this includes the substantial amount of noise and systematic variations between spectra caused by sample degradation over time, ionization suppression and other parameters that can be reduced.

Baseline adjustment is important because it reduces background noise, and a drifting baseline introduces serious distortion of ion intensities. Normalization is used to remove effects from systematic variations among spectra due to variations in amounts of protein or variation in the detector sensitivity. By smoothing the raw spectra, we can reduce the effect of some mass-per-charge (m/z) values that appear as peaks but which are hard to verify by independent experiments. When cells become cancerous they can release unique proteins and other molecules into the blood and other body fluids, and these molecules may serve as early biomarkers or indicators of cancer. Peak detection deals with the selection of m/z values that display a reasonable intensity compared with those that display noise. Clustering is performed by using a clustering algorithm that aligns a peak with slightly different m/z values caused by noise.

The preprocessed data is then passed through an efficient feature selection algorithm [26-27] for biomarker identification. In this thesis, the main focus is on tuning an advanced feature selection algorithm [26-27] for biomarker identification. The central hypothesis is that the proposed feature selection algorithm can identify a compact set of robust biomarkers, and can provide some advantages over existing techniques. The proposed algorithm evaluates peak combinations by considering their interactions. Correlated peaks will be eliminated automatically. The algorithm produces a list of near-optimal combinations for all possible number of peaks with sensitivity and specificity calculated for each of the combinations. Users can then choose a combination based on its sensitivity and specificity.

1.7 Motivation

Building computational models for biomarker identification is important because the output of high-resolution mass spectrometry is a large dataset, and the analysis strategy must face a number of technical challenges. Based on recent projects by EDRN, an efficient feature selection algorithm is proposed and applied to mass spectrometry data from a prostate cancer tissue sample to see if it can improve the performance in identifying the cancer specific biomarkers.

1.8 Goal

The goal of this thesis is to show that computational methods can be useful in narrowing the protein biomarker candidates. In this thesis, biomarker identification is achieved by applying the proposed three-step pipeline to the high dimensional raw mass

spectrometry data: (1) data preprocessing to preprocess and reduce the dimensionality of each spectrum, (2) feature selection to select the discriminating pattern, and (3) classification to classify each spectrum as cancer or non-cancer based on the identified pattern (biomarker).

1.9 Thesis Outline

Chapter 2 reviews various other feature selection algorithms implemented for comparison with the proposed algorithm.

Chapter 3 presents the proposed pipeline for detection of cancer specific biomarkers, including three steps: (i) data preprocessing, (ii) the feature selection algorithm, and (iii) classification. Various comparison metrics are also discussed.

Chapter 4 presents experimental results.

Chapter 5 provides conclusions and directions for future work.

CHAPTER 2

RELATED WORK

Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. All the feature selection algorithms used for the comparison with the proposed feature selection algorithm are described in this section.

2.1 AUC Score

The receiver operating characteristic curve (ROC) is commonly used to measure the performance of a diagnostic system in terms of its "hit-or-miss" behavior. By computing the ROC curve for each feature individually, one can determine the ability of that feature to separate samples into correct groups. Measuring the area under the ROC curve (AUC) [28] then gives an indication of the feature's probability of being a successful biomarker. The AUC score for a given feature is then obtained by integrating over the ROC curve for that feature. Higher AUC scores signify better feature candidates.

2.2 J5 Test

The J5 test [29] is a gene-specific ratio between the mean difference in expression intensity between two groups, A and B, to the average mean group difference of all m genes.

$$J5_i = \frac{\overline{A}_i - \overline{B}_i}{\frac{1}{m}\sum_{j=1}^{m}\left|\overline{A}_j - \overline{B}_j\right|} \tag{2-1}$$

The J5 test is likely to be useful in pilot studies where, due to high variance, t-tests are likely to exhibit unacceptably low specificity (high false discovery rates).

2.3 minimum-Redundancy-Maximum-Relevance (mRMR)

Selecting features that correlate the strongest to a target variable has been called the "maximum relevance selection." On the other hand, features can be selected to be different from each other, while they still have high correlation to the target variable. This scheme, called "minimum-Redundancy-Maximum-Relevance" selection, has been found to be more powerful than the maximum relevance selection [30].

As a special case, statistical dependence between variables can be used instead of correlation. Mutual information can be used to quantify the dependency. In this case, it can be shown that minimum-Redundancy-Maximum-Relevance (mRMR) is an approximation to maximize the dependency between joint distribution of the selected features and the target variable.

2.4 Random Search

Feature selection can also be reinforced by a learning algorithm; this approach is usually referred to as a wrapper selection method. A randomized search for feature selection generates random subsets of features and assesses their quality independently with the learning algorithm. Later, it selects a pool of the most frequent good features. Li *et al.* in [31] applied this concept to the analysis of protein expression patterns.

2.5 Genetic Algorithm Search

Genetic algorithms optimize search results for problems with large data sets. Genetic algorithms have been applied to phylogenetic tree building, gene expression and mass spectrometry data analysis, and many other areas of Bioinformatics that have large and computationally expensive problems [23].

A genetic algorithm requires an objective function, also known as the fitness function that describes the performance of a feature or a feature subset. The genetic algorithm tests candidate features using the fitness function and then determines which one gets passed on to or removed from each subsequent generation. The fitness function is usually optimized by several steps, including crossover, mutation, reproduction and fitness evaluation.

CHAPTER 3

MATERIALS AND METHODS

3.1 Materials

In this thesis, MALDI-MS tissue imaging data from prostate cancer tissue collected at Eastern Virginia Medical School (EVMS) is used. Protein profiles of a tumor and surrounding tissue will be used as inputs to the feature selection algorithm. There are 974 spectra, 27 of which belong to cancer, and the rest belong to normal spectra in the tissue sample. The dimension of each spectrum is 82,756. Three of the spectra were illustrated in Figure 3.1.

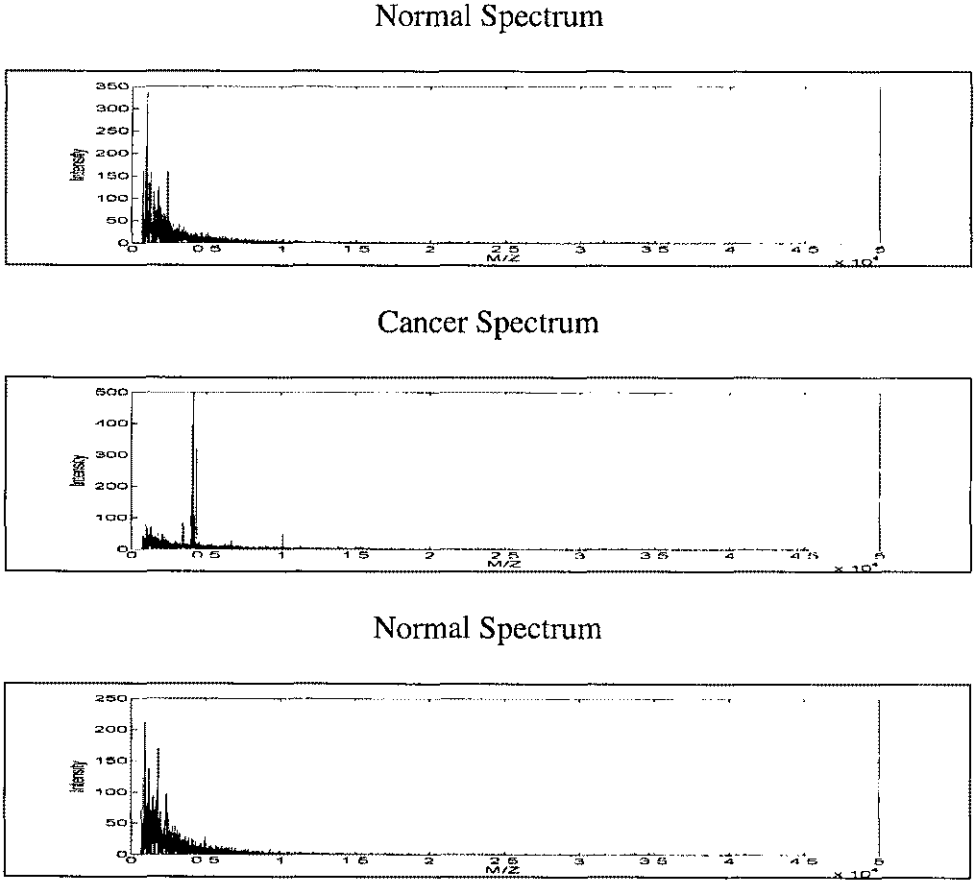

Figure 3.1: Individual spectra plots.

3.2 Methods

3.2.1 Preprocessing Techniques

The data preprocessing techniques [20] that are applied to the raw mass spectrum data are explained as follows.

3.2.1.1 Baseline Adjustment

This step is required to remove the ion overload and chemical noise that are usually higher at smaller *m/z* values. There is no general solution to this problem because baseline characteristics vary from one experiment to another and each spectrum has to be assessed individually. For the mass spectrum data considered, the MATLAB function Yout = msbackadj (MZ, Y) is used, which estimates the baseline within multiple shifted windows of width 200 *m/z*. It regresses the varying baseline to the window points using a spline approximation and adjusts the baseline of the spectrum Y to Yout. The result of baseline adjustment is shown in Figure 3.2 and its zoom-in is shown in Figure 3.3. The blue lines show the original spectra, and the regressed baseline is shown in red. The black cross marks are the estimated baseline points.

3.2.1.2 Smoothing

Smoothing is done to denoise and thus enhance the signal-to-noise ratio by using wavelets [32]. The *m/z* values lower than 3,000 due to large noise and *m/z* values greater than 10,000 due to low intensities are discarded.

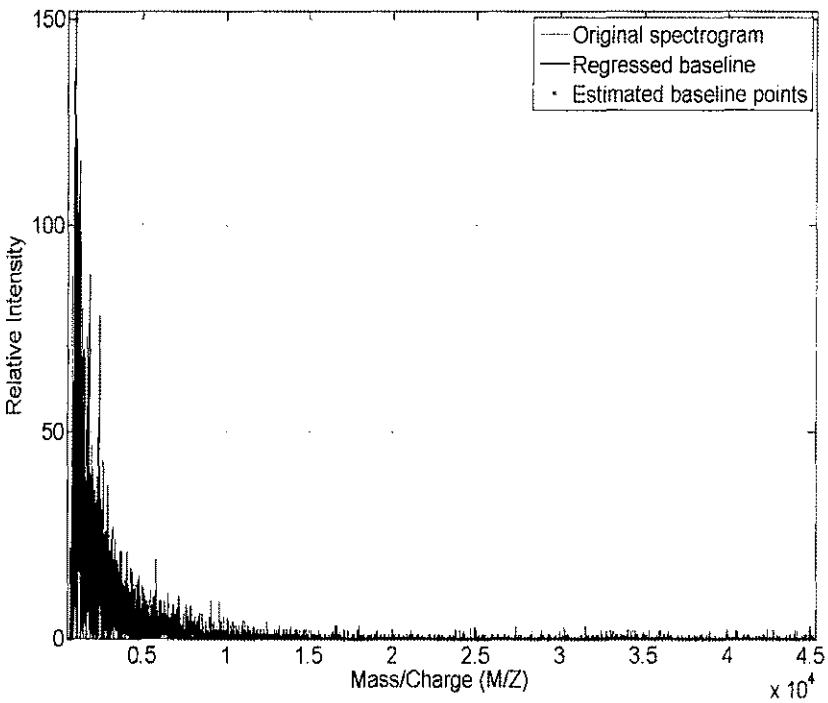

Figure 3.2: Result of baseline adjustment.

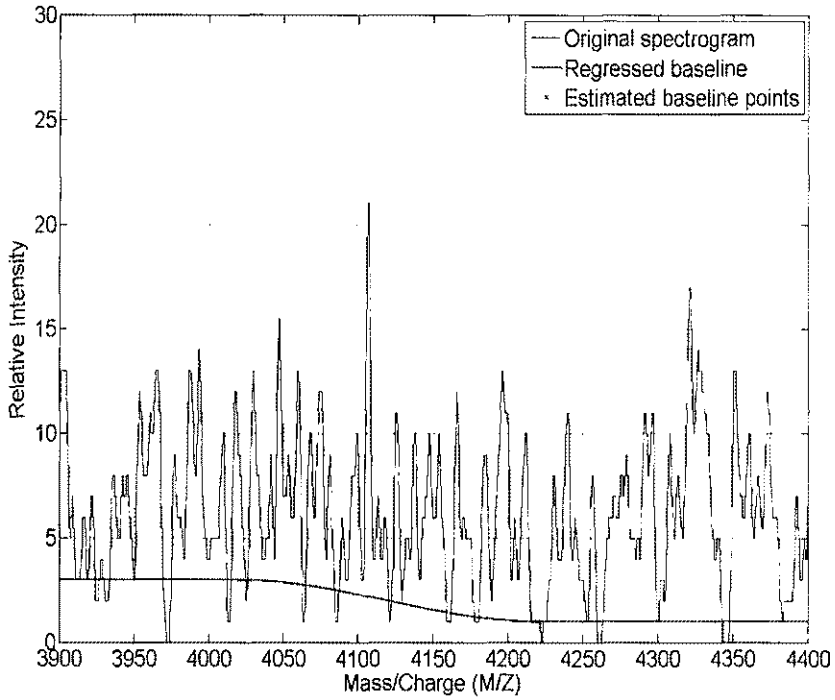

Figure 3.3: Zoom-in of Figure 3.2.

3.2.1.3 Normalization

When dealing with multiple spectra it is a good practice to remove effects from systematic variation among spectra due to varying amounts of protein or variation in detector sensitivity. A global normalization procedure [33] where mass intensities for the same peak from different spectra are scaled (divided) by a common factor is used. For a given peak the area under the peak is computed, i.e., the sum of all intensities for this peak from all spectra. The common factor for each peak is then defined as the ratio of the area under this peak to the median of areas of all the other peaks in a single spectrum.

3.2.1.4 Peak Detection

A crucial step for the identification and quantification of proteins in mass spectra is to find *m/z* values that correspond to high peak intensities [34]. Peak detection deals with the selection of mass points with reasonable intensity and S/N ratio. The peak detection method satisfies the criteria; the intensity should exceed a specified threshold value of 10, below which all intensity values in the spectrum are zeroed. After smoothing and peak detection are performed, a total of 75,719 peaks from the available 974 spectra are obtained. Each spectrum represents the protein profile of one spot (cells) in the prostate tissue sample. The result after smoothing, normalization and peak detection for one spectrum is shown in Figure 3.4 and its zoom-in is shown in Figure 3.5. The original spectrum in blue is unclear because of noise. The denoised spectrum is shown in green, and the detected peaks are shown in red cross marks. The steps mentioned above greatly reduce the unnecessary peaks and make the task of biomarker identification relatively simple.

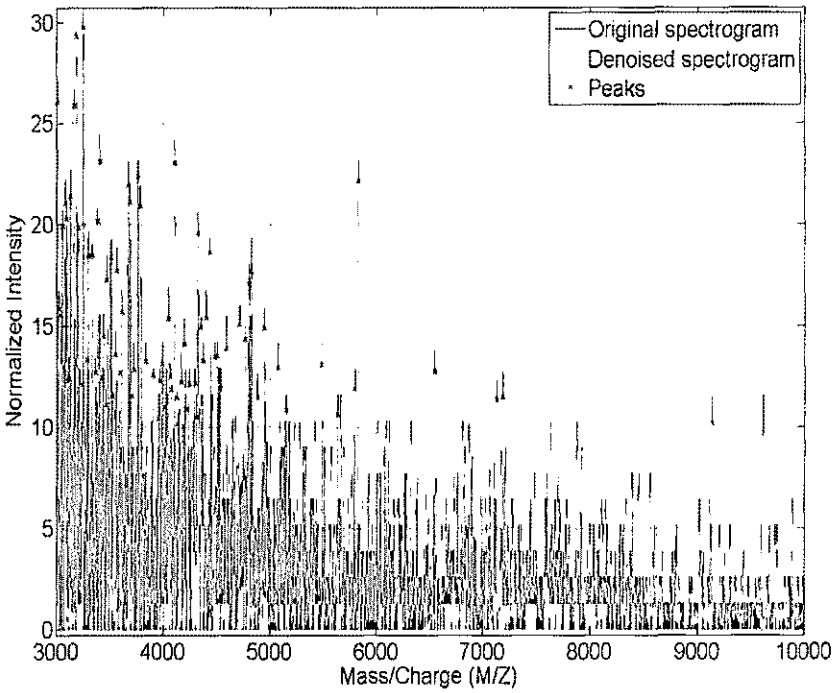

Figure 3.4: Result after smoothing, normalization and peak detection.

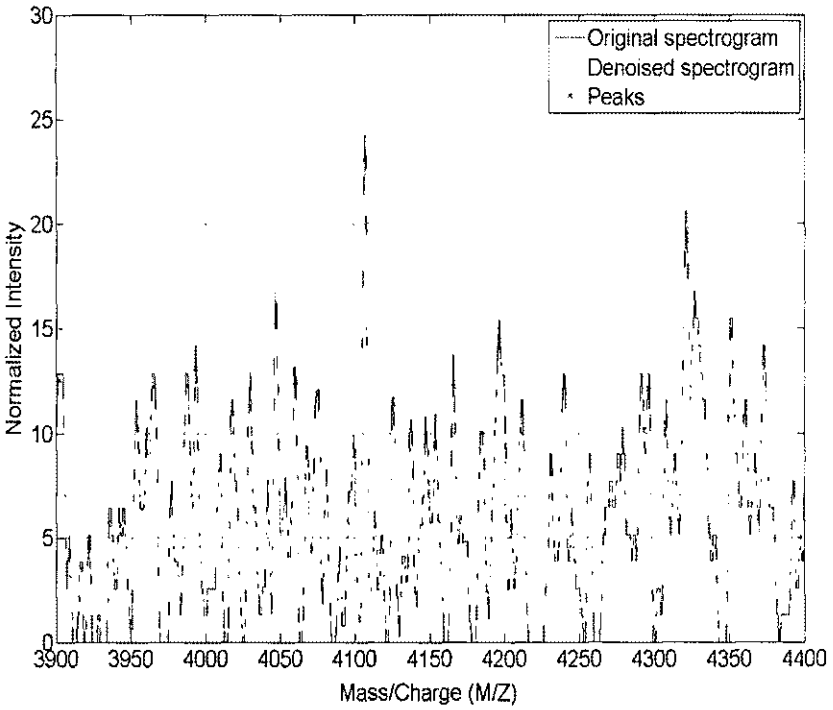

Figure 3.5: Zoom-in of Figure 3.4.

3.2.1.5 Clustering

This is the crucial step in the identification of cancer specific biomarkers. The first step in this process is to align or 'cluster' the same protein from different spectra, that is, to assign a cluster number to every protein found in all spectra even if they have slightly different *m/z* values due to noise. All peaks detected from the available 974 spectra are projected onto a single axis and clustered [25]. The result of projection onto a single axis is shown in Figure 3.6, and its zoom-in is shown in Figure 3.7.

Many of the peaks represent the same protein and are not aligned due to the fact that the mass spectra exhibit shifts along the horizontal axis between multiple spectra and the instruments have a small error on the *m/z* scale. Thus, detected peaks that have masses within the percentage range are considered identical. We merged peaks that have *m/z* measurements within 0.13% of each other and assigned the new peak the average of *m/z* values. The 0.13% threshold is selected because in this range there is no more than one peak from the individual spectra that contributes to a different protein. It also yielded better results when compared to different threshold values.

After this step we got a total of 820 clusters which represented different proteins. The 820 protein clusters are shown in Figure 3.8, and their zoom-in is shown in Figure 3.9. The next step is to back project these peaks onto individual spectra and use them to identify biomarkers. In the process of back projection if there is a peak in one individual spectrum corresponding to a cluster, the intensity of the peak is kept as is. If there is no peak then the cluster point is replaced by zero. The result of back projection is shown in Figure 3.10, and its zoom-in is shown in Figure 3.11. The top and the bottom spectra belong to normal spots, and the middle spectrum belongs to a cancer spot.

After the preprocessing step, the peaks obtained are denoted as $\{x_p, i_p\}_{p=1}^{N_v}$, where $x_p \in R^N$ and $i_p \in R$, x_p is a vector containing all peaks detected from p^{th} spectrum and i_p is a class ID (1: normal, 2: cancer) associated with this spectrum. N (820) is the total number of peaks detected from all spectra, and N_v (974) is the total number of spectra. The class ID is obtained from the pathological analysis results and is considered to be ground truth. Those peaks then go through the feature selection algorithm described in the next section.

3.2.2 Biomarker Identification

The purpose of this analysis is to identify optimal *m/z* values or candidate biomarkers from the preprocessed mass spectral data that can discriminate normal from cancer spots. Methods are usually tailored towards classifying spectra into nominal categories based on a set of peaks detected from the spectra. There are usually hundreds of peaks detected from a set of mass spectra. Using all of the detected peaks for the classification can have side effects including the curse of dimensionality, convergence difficulties and large validation errors. In order to avoid these problems, feature selection is usually utilized to generate a compact subset of peaks that leads to accurate models for the available data.

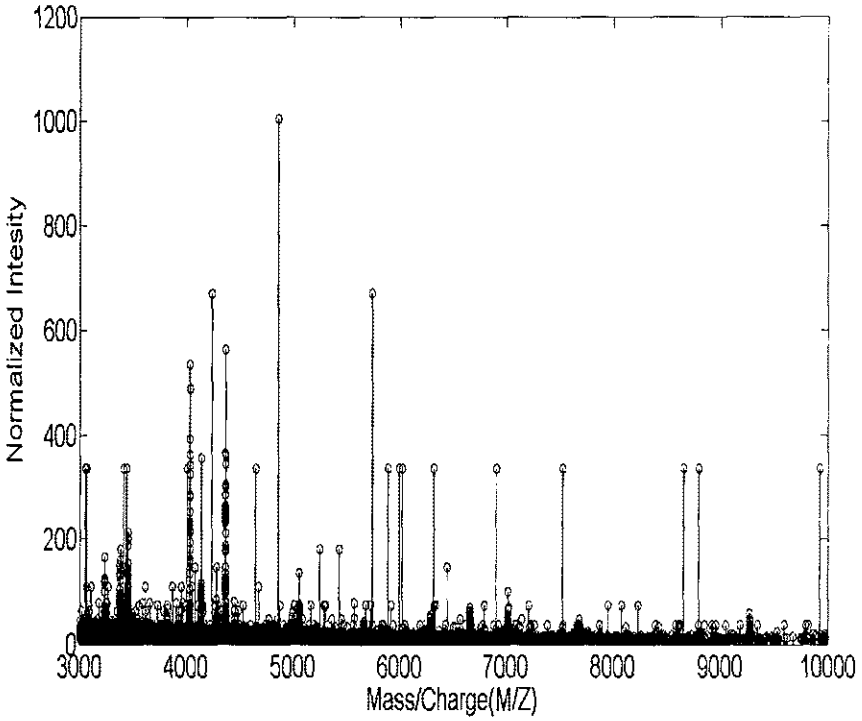

Figure 3.6: Result of projection of all the 974 spectra onto a single axis.

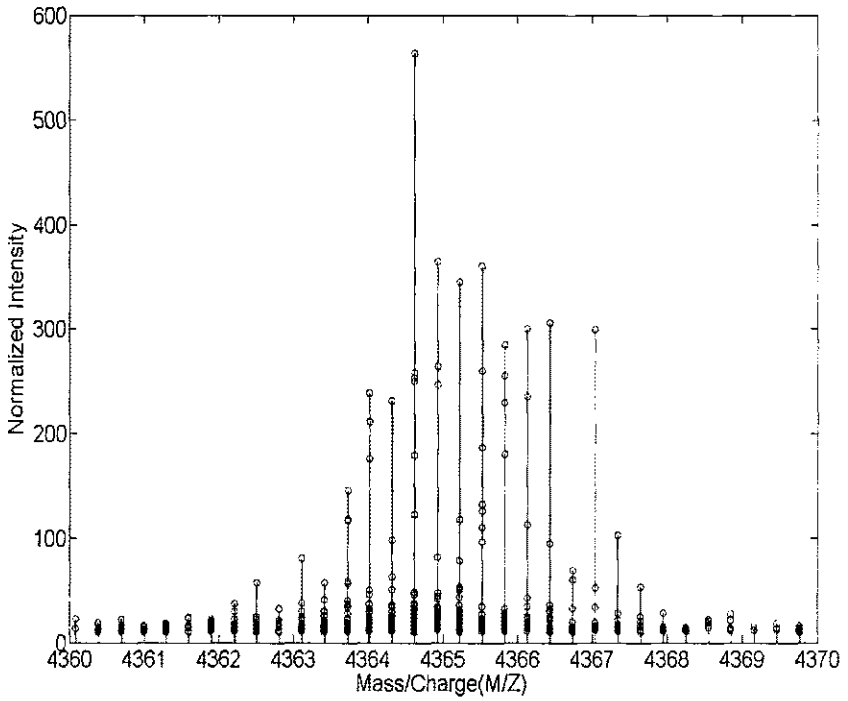

Figure 3.7: Zoom-in of Figure 3.6.

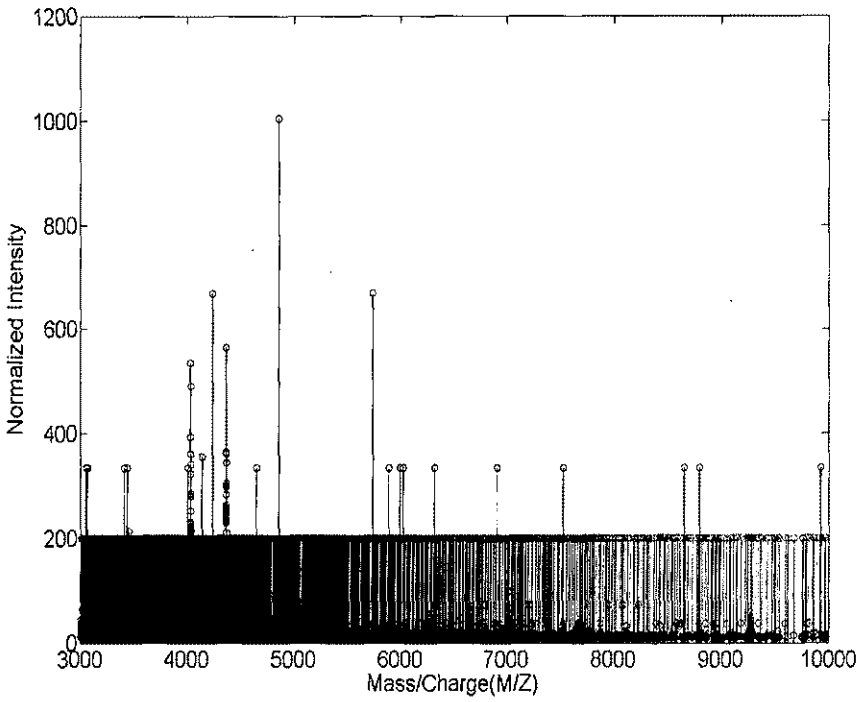

Figure 3.8: Result of clustering (red lines represent the 820 clusters).

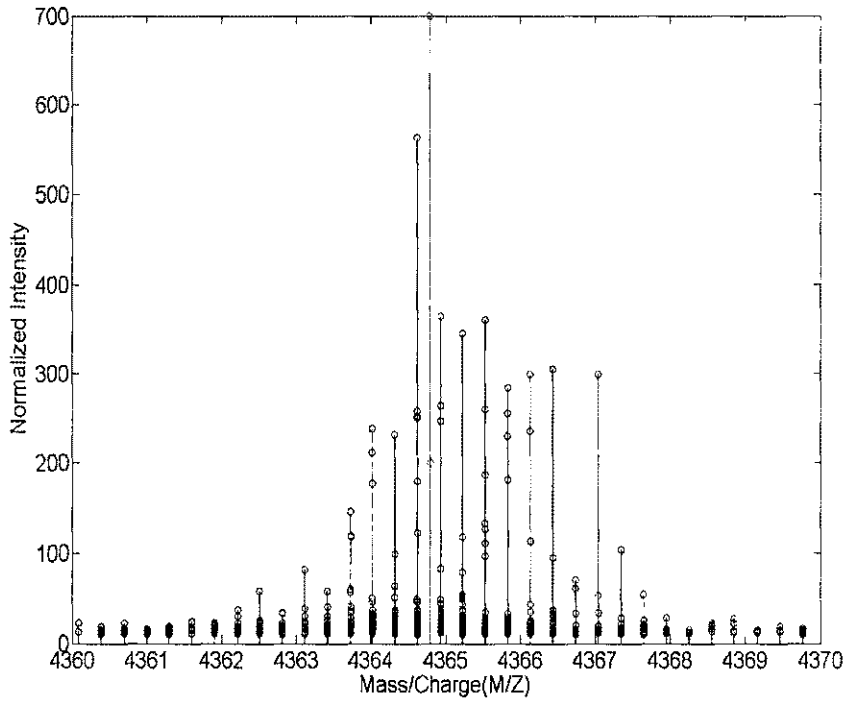

Figure 3.9: Zoom-in of Figure 3.8.

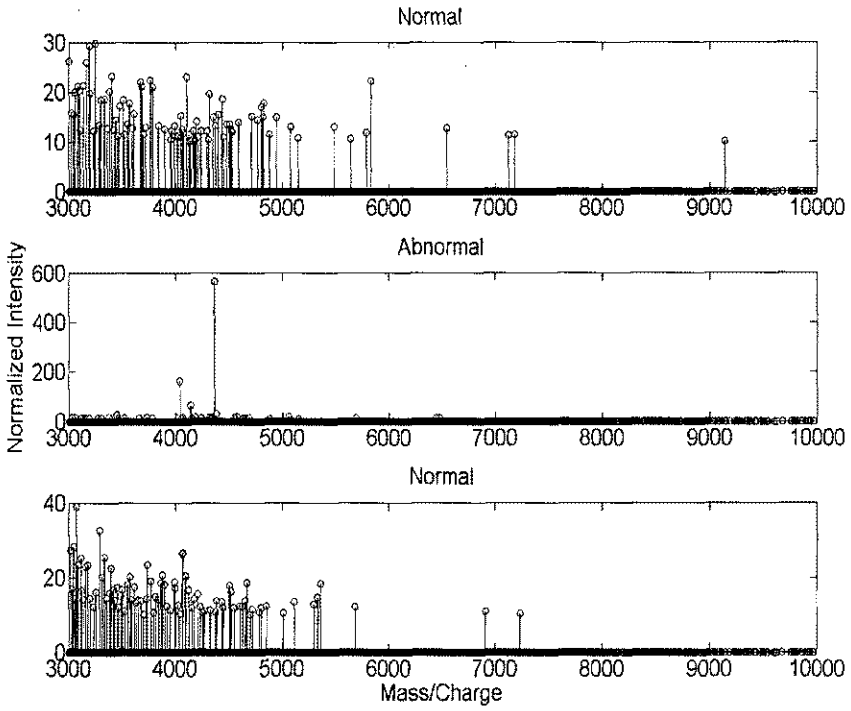

Figure 3.10: Result of the back projection.

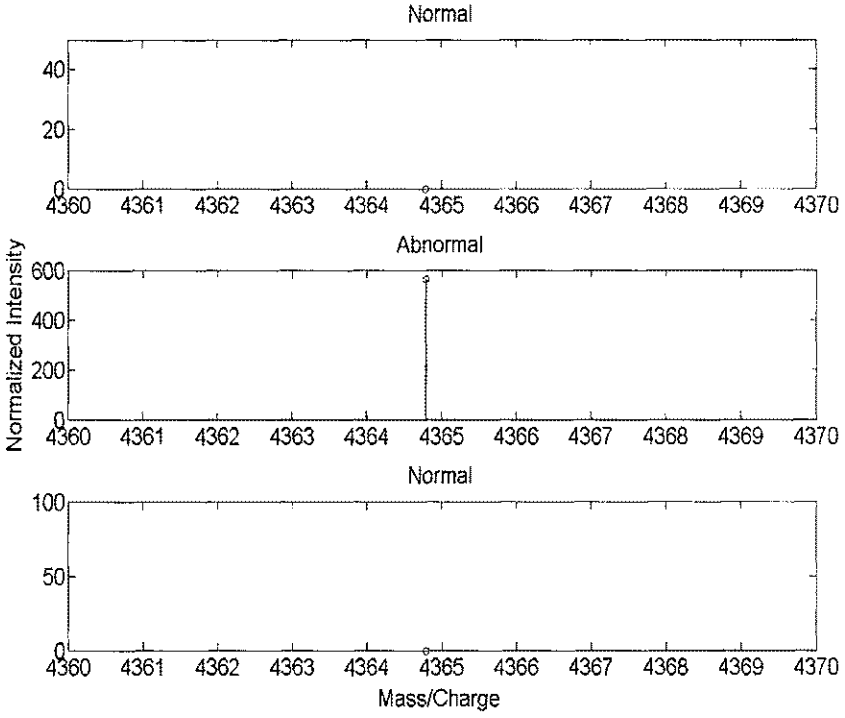

Figure: 3.11: Zoom-in of Figure 3.10.

The feature selection algorithm for identification of cancer specific biomarkers proposed in this thesis is discussed in detail in this section. The algorithm consists of three important components: a piecewise linear classifier, an output reset algorithm and a floating search algorithm.

3.2.2.1 Piecewise Linear Classifier

Neural classifiers including the piecewise linear classifier (PLC) are usually designed by minimizing the standard training error given by the formula [35],

$$E = \sum_{i=1}^{N_{class}} E(i) = \frac{1}{N_v} \sum_{i=1}^{N_{class}} \sum_{p=1}^{N_v} [t_p(i) - y_p(i)]^2 \tag{3-1}$$

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

$$y_p(i) = \sum_{j=1}^{N+1} w^{(q)}(i, j) x_p^{(q)}(j) \tag{3-2}$$

where N is the number of features, $w^{(q)}(i, j)$ denotes the model weight to the i^{th} output from the j^{th} feature in the q^{th} cluster, $x_p^{(q)}(j)$ is the j^{th} feature in the q^{th} cluster, and $x_p^{(q)}(N+1)$ is the bias term that equals one.

The PLC approximates the general Bayes discriminant [35]. The available data is divided into a set of clusters where a local linear model is obtained for each cluster, by solving sets 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

$$i_c = \arg\max_i y_p(i), \tag{3-3}$$

we say that PLC classified the current pattern correctly. Otherwise, a classification error is counted.

Note that the error function E in equation (3-1) is too restrictive in at least two ways. First, if each individual output vector has a different constant bias added to it, E could be increased or decreased, with no effect on the classification error. Second, if an output has the correct sign but a magnitude larger than 1, E will increase while the classification error will be unaffected or decrease. In order to take advantage of these effects, an Output Reset (OR) algorithm is developed to relax the restrictions.

3.2.2.2 The OR Algorithm

In the OR algorithm [36] a number a_p is first added to each desired output vector, which compensates for unwanted bias in the actual output vector y_p. Second, a function $d_p(i)$ is added to the desired output vector to compensate E for errors calculated when the output has the correct sign but is larger than 1 in magnitude. The error function E in equation (3-1) becomes,

$$E' = \frac{1}{N_v}\sum_{p=1}^{N_v}\sum_{i=1}^{N_{class}}[t_p'(i) - y_p(i)]^2 \tag{3-4}$$

where

$$t_p'(i) = t_p(i) + a_p + d_p(i) \tag{3-5}$$

and where $d_p(i)$ is a function of p and i to be defined later. The goal is to find $a_p, d_p(i)$ and $y_p(i)$ that minimize E', under the following two conditions:

1. The difference $\left|t_p'(i_c) - t_p'(i_d)\right|$ must be larger than or equal to 2. Without this condition, E' can be minimized by setting the network weights and the difference $\left|t_p'(i_c) - t_p'(i_d)\right|$ to zero.
2. Each change made to $a_p, d_p(i)$ and $t_p'(i)$ (through changes in the network weights) must reduce E' or keep it unchanged.

Using these two conditions the following methods can be used for computing a_p and $d_p(i)$.

Method 1. Determination of a_p: In order to minimize E' with respect to a_p, it is sufficient that the first derivative of E' with respect to a_p be zero, yielding

$$a_p = \frac{1}{N_{class}} \sum_{i=1}^{N_{class}} [y_p(i) - t_p(i) - d_p(i)] \tag{3-6}$$

Method 2. Determination of $d_p(i)$: Ignore condition (1), $d_p(i)$ can be found such that the term $\left[t_p(i) + a_p + d_p(i) - y_p(i)\right]^2$ is zero, yielding $d_p(i) = y_p(i) - t_p(i) - a_p$, which satisfies condition 2. However, in order to satisfy condition 1, $d_p(i)$ is modified such that $d_p(i_c) \geq 0, d_p(i_d) \leq 0$. In summary,

a. If $y_p(i_c) \geq t_p(i_c) + a_p$, then choose $d_p(i_c) = y_p(i_c) - t_p(i_c) - a_p$

b. If $y_p(i_d) \le t_p(i_d) + a_p$, then choose $d_p(i_d) = y_p(i_d) - t_p(i_d) - a_p$

c. Otherwise, choose $d_p(i_c) = 0$ and $d_p(i_d) = 0$

Method 1 is used initially to find a_p with $d_p(i)$ set to zero. Method 1 and Method 2 are alternatively performed for three iterations, and $t_p(i)$ is replaced by $t_p'(i)$. It is found that three iterations are sufficient for a_p and $d_p(i)$ to converge. Note that both conditions are satisfied in the two methods. In summary, traditional mean square error (MSE)-type training attempts to force all training patterns to be support vectors. This is remedied by using OR. The OR algorithm is extremely useful for peak selection because the detected peaks are highly unbalanced, i.e., there are many more normal spectra than cancel spectra. The OR algorithm does not consider the spectra that is far away from the decision boundary. Hence, there are only portions of the spectra utilized in the peak selection process.

3.2.2.3 Floating Search Algorithm

The floating search method [27] is designed through the Piecewise Linear Orthonormal Least Square (PLOLS) procedure in this section. The PLOLS procedure utilizes the modified Schmidt procedure to make each feature in each cluster orthonormal. This procedure passes through the data set once, and all information needed for searching a good combination of features is stored in the auto-correlation and cross-correlation matrices. Therefore, this feature selection algorithm is very efficient as only one data pass is required.

Based on equations (3-1) and (3-2), the modified desired output may be represented in matrix form as,

$$t^{'} = x^{(q)} w^{(q)} + \Xi^{(q)} \tag{3-7}$$

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

$$t^{'} = \Theta^{(q)} A^{(q)} w^{(q)} + \Xi^{(q)} = \Theta^{(q)} w_o^{(q)} + \Xi^{(q)} \tag{3-8}$$

Based on the following four definitions, our proposed feature selection algorithm is described as follows. Let $X(d) = \{x(i) : 1 \le i \le d, x(i) \in Z\}$ be a subset of d features from the set $Z = \{z(i) : 1 \le i \le N\}$ of N available features and M outputs. Suppose the feature space is partitioned into N_c clusters and obtained the PLO system as equation (3-8).

Definition 1: The individual fitness of one feature, $x(i)$, is

$$S_0(x(i)) = \sum_{k=1}^{M} \sum_{q=1}^{N_c} (w_o^{(q)}(k,i))^2, \tag{3-9}$$

which is the total variance explained for all outputs due to the i^{th} feature.

Definition 2: The fitness of a set of d features $X(d)$, is measured as

$$J(X(d)) = \sum_{i=1}^{d} \sum_{k=1}^{M} \sum_{q=1}^{N_c} (w_o^{(q)}(k,i))^2, \tag{3-10}$$

which is the total variance explained for all outputs due to all features in the set $X(d)$.

Definition 3: The fitness $S_{d-1}(x(i))$ of the feature $x(i), 1 \le i \le d,$ in the set $X(d)$ is defined by

$$S_{d-1}(x(i)) = \sum_{k=1}^{M} \sum_{q=1}^{N_c} (w_o^{(q)}(k,i))^2, \tag{3-11}$$

where $x(i)$ is the last feature in the set $X(d)$ that is made orthonormal to the other bases in the modified Schmidt procedure.

Definition 4: The fitness $S_{d+1}(x(i))$ of the feature $x(i)$ with respect to $X(d)$, where $x(i) \in Z - X(d)$, is

$$S_{d+1}(x(i)) = \sum_{k=1}^{M} \sum_{q=1}^{N_c} (w_o^{(q)}(k,i))^2, \tag{3-12}$$

where $x(i)$ is made orthonormal to $X(d)$, to get $w_o(k,i), k = 1,2,...M$. These four definitions are fitness measures for one feature or feature combinations that will guide the feature selection process.

3.2.2.4 Algorithm Description

The proposed feature selection algorithm for selecting N_s features from N available features is described as follows.

1. Using the trail and error method, an appropriate number of clusters, N_c, is determined, which will be used in the PLC.
2. Design a N_c cluster PLC for the data by solving a set of linear equations for each cluster.
3. Change desired output using the OR algorithm.
4. Based on the above four definitions, search a list of good feature combinations using the floating search algorithm [27].

3.2.2.5 Advantages of Proposed Algorithm

Advantages of the proposed algorithm are as follows:

1. It selects features rather than a combination of all the features such as those selected by transformation based methods (PCA, Wavelet).
2. It considers interactions among features and measures the correlations via the amount of explained variance by features.
3. It is computationally efficient.
4. It automatically handles the extremely unbalanced data sets where the number of instances in some classes is significantly more than those in other classes.
5. The algorithm produces a list of best combinations that contain different numbers of features; users then have the flexibility to choose one based on performance.

3.2.3 Classification

In this thesis, the multi layer perceptron (MLP) classifier is used for the classification task. After the compact sets of features are selected from the previous step, an MLP classifier is used to classify the spectra to one of the two classes (normal or cancer) [36-37]. Since the class label of each training sample is provided, this step is also known as supervised learning (i.e., the learning of the model is 'supervised' in that it is told to which class each training sample belongs) [38].

The classifier utilized a new objective function that had more free parameters than the classical objective functions and used an iterative minimization technique to solve multiple sets of numerically ill-conditioned linear equations. An enhanced feedforward network training algorithm was also used to reduce a separate error function

with respect to hidden layer weights. The MLP classifier is explained in detail in the following sections.

3.2.3.1 OR Enhanced MLP Training

In this section, the integration of OR into more advanced MLP training algorithms is discussed; all the weights are subjected to optimization. There are many well-developed training algorithms, including the Back Propagation (BP), Conjugate Gradient (CG) and Levenberg-Marquardt (LM) algorithms. Training error can be further decreased when OR is used in most algorithms.

In this thesis an algorithm called Output Weight Optimization-Hidden Weight Optimization (OWO-HWO) [39] is utilized, which can be used in the training of feed-forward neural networks such as the MLP. In OWO-HWO, output weights and hidden unit weights are alternately modified to reduce the training error. Since the output units have linear activation functions, in this method the OWO procedure is used to obtain output weights by solving linear equations, whereas the HWO is utilized to calculate the hidden weight changes by minimizing a mean-square error between the desired and the actual net function.

3.2.3.2 Review of Output Weight Optimization

Applying OWO to the three layer, fully connected MLP, basis functions are defined as

$$X_p(k) = \begin{Bmatrix} x_p(k) & & 1 \le k \le N \\ 1 & \text{for} & k = N+1 \\ O_p(k-N-1) & & N+2 \le k \le N+N_h+1 \end{Bmatrix} \tag{3-13}$$

where $O_p(j)$ is the j^{th} hidden unit output activation for the p^{th} pattern. $O_p(N+1)=1$ to handle the hidden unit and output unit biases. Substituting equation (3-2) into $E(i)$ of equation (3-1), the mean-squared error for the i^{th} output can be rewritten as

$$E(i)=\frac{1}{N_v}\sum_{p=1}^{N_v}\left[t_p(i)-\sum_{k=1}^{N_u}w_o(i,k)X_p(k)\right]^2 \tag{3-14}$$

where $N_u=N+N_h+1$. Taking the gradient $E(i)$ with respect to the output weights gives

$$g(m)=\frac{\partial E(i)}{\partial w_o(i,m)}=-2\left[c(i,m)-\sum_{k=1}^{N_u}w_o(i,k)\cdot r(k,m)\right] \tag{3-15}$$

where $1\le m\le N_u$. The cross-correlation $c(i,m)$ and auto-correlation $r(k,m)$ are defined as

$$c(i,m)=\sum_{p=1}^{N_v}t_p(i)\cdot X_p(m) \tag{3-16}$$

$$r(k,m)=\sum_{p=1}^{N_v}X_p(k)\cdot X_p(m) \tag{3-17}$$

Setting $g(m)$ to zero, we get

$$\sum_{k=1}^{N_u}w_o(i,k)\cdot r(k,m)=c(i,m)\qquad 1\le m\le N_u \tag{3-18}$$

Each value of i has a set of N_u equations in N_u unknowns. Since those linear equations generally are ill-conditioned, the conjugate gradient approach can be utilized to get the output weights that minimize $E(i)$.

OWO is only adequate for generating a useful initial network after the hidden weights have been initialized. Note that the hidden weights are not updated in OWO training.

3.2.3.3 Review of OWO-HWO

In OWO-HWO [39] initially a set of N_v training patterns (x_p, t_p) where the p^{th} input vector x_p and p^{th} desired output vector t_p having a dimension N and N_{class}, respectively, are given. A three layer, fully connected MLP network with sigmoid activation function for the hidden layer is used. For the p^{th} pattern, the j^{th} hidden unit net and activation functions are

$$net_p(j) = \sum_{k=1}^{N+1} w(j,k) \cdot x_p(k) \tag{3-19}$$

$$O_p(j) = f(net_p(j)) = \frac{1}{1+\exp(-net_p(j))} \tag{3-20}$$

the i^{th} observed output is

$$y_p(i) = \sum_{k=1}^{N+1} w_{oi}(i,k) \cdot x_p(k) + \sum_{j=1}^{N_h} w_{oh}(i,j) \cdot O_p(j) \tag{3-21}$$

where $w_{oi}(i,k)$ and $w_{oh}(i,j)$ are weights connecting to the i^{th} output unit from the k^{th} input and j^{th} hidden unit, respectively. The output weighs $w_{oi}(i,k)$ and $w_{oh}(i,j)$ can be found using the OWO method. In the HWO procedure, the hidden weights $w(j,k)$ are updated by minimizing a separate error function for each hidden unit. For the j^{th} hidden unit and p^{th} pattern, the desired net function $net_{pd}(j)$ is constructed as [47]

$$net_{pd}(j) \cong net_p(j) + Z \cdot \delta_p(j) \tag{3-22}$$

Z is the learning rate, and $\delta_p(j)$ is the delta function of the j^{th} hidden unit and is defined as

$$\delta_p(j) = f'(net_p(j)) \sum_{i=1}^{N_c} \delta_{po}(i) w_o(i,j) \tag{3-23}$$

where $\delta_{po}(i)$ is the delta function of the i^{th} output layer,

$$\delta_{po}(i) = t_p(i) - y_p(i) \tag{3-24}$$

The hidden weights are updated as

$$w(j,k) \leftarrow w(j,k) + Z \cdot e(j,k) \tag{3-25}$$

where $e(j,k)$ is the hidden weight change. With the basic operations and equations (3-22) to (3-24), the following equation is used to solve for changes in the hidden weights,

$$net_{pd}(j) + Z \cdot \delta_p(j) \cong \sum_{k=1}^{N+1} \left[w(j,k) + Z \cdot e(j,k) \right] \cdot x_p(k) \tag{3-26}$$

and obtained

$$\delta_p(j) \cong \sum_{k=1}^{N+1} e(j,k) \cdot x_p(k) \tag{3-27}$$

Before solving equation (3-27) in the least square sense, an objective function [48] for the j^{th} hidden unit is defined as

$$E_\delta(j) = \sum_{p=1}^{N_v} \left[\delta_p(j) - \sum_{k=1}^{N+1} e(j,k) x_p(k) \right]^2 f'(net_p(j)) \tag{3-28}$$

which is minimized with respect to $e(j,i)$ using the conjugate gradient method, and the hidden weights change $e(j,k)$ is obtained; the hidden weights is updated by performing equation (3-25).

3.2.3.4 Algorithm Description for OR Combined MLP Training

Using both OR and OWO-HWO, the following algorithm is constructed,

1. Initialize all the weights and thresholds as small random numbers in the usual manner. Pick a value for the maximum number of iterations, N_{it}. Set the iteration number to 0.
2. Increment i_t by 1. Stop if $i_t > N_{it}$.
3. For each input vector, calculate the hidden unit outputs $O_p(j)$. If $i_t = 1$, i.e., in the first iteration, accumulate the cross- and auto-correlation matrices $c(i,m)$ and $r(k,m)$ as in equation (3-16) and (3-17). Otherwise, if $i_t > 1$, use the OR algorithm to change the desired output $t_p(i)$ to $t_p^{'}(i)$ for each pattern and accumulate the cross-correlation $c(i,m)$ as in equation (3-16) with $t_p(i)$ replaced by $t_p^{'}(i)$.
4. Using OWO, solve linear equations for the output weights $w_{oi}(i,k)$ and $w_{oh}(j,k)$, and calculate E.
5. If E decreases, go to Step 8. If E increases, modify Z as $Z = 0.5 \cdot Z$, reload the previous best hidden weights and go to Step 8.
6. Make a second pass through the training data. Calculate the hidden weight changes using HWO with $t_p^{'}(i)$ in place of $t_p(i)$.
7. Calculate the learning factor λ using the method described by Magoulas [40].
8. Update the hidden unit weights as in equation (3-25).
9. Go to Step 2.

3.2.4 Comparison Metrics

For the comparison of different feature selection algorithms, the below mentioned techniques and metrics are employed.

3.2.4.1 Cross Validation (CV)

In this thesis 5-fold cross validation is used. The data set is divided into 5 subsets, and each time one of the 5 subsets is used as the test set the other 4 subsets are put together to form a training set. The process is repeated 5 times so that each subset is used for testing once. After the 5-fold CV the 5 test results are pooled together to compute sensitivity and specificity.

3.2.4.2 Sensitivity and Specificity

Sensitivity and specificity are statistical measures of the performance of a binary classification problem. The sensitivity or the recall rate measures the proportion of actual positives that are correctly identified as such (i.e. the percentage of cancer spots that are identified as having the condition); the specificity measures the proportion of negatives that are correctly identified (i.e. the percentage of non-cancer spots that are identified as not having the condition).

A sensitivity of 100% means that the test recognizes all cancer spots. Sensitivity alone does not tell us how well the test predicts other classes (that is, about the negative cases). In the binary classification, as illustrated above, this is the corresponding specificity test or, equivalently, the sensitivity for the other classes.

$$sensitivity = \frac{numberofTruePositives}{numberofTruePositives + numberofFalseNegatives} \quad (3\text{-}29)$$

A specificity of 100% means that the test recognizes all normal spots as normal. The maximum is trivially achieved by a test that claims every spot as normal regardless of the true condition. Therefore, the specificity alone does not tell us how well the test recognizes positive cases. We also need to know the sensitivity of the test to the class or, equivalently, the specificities to the other classes.

$$specificity = \frac{numberofTrueNegatives}{numberofTrueNegatives + numberofFalsePostives} \quad (3\text{-}30)$$

3.2.4.3 ROC curve

A ROC curve is a graphical representation of the trade off between the false negative and false positive rates. Equivalently, the ROC curve [35] is the representation of the tradeoffs between sensitivity and specificity. By tradition, the plot shows the false positive rate on the X axis and 1 - the false negative rate on the Y axis. You could also describe this as a plot with 1-specificity on the X axis and sensitivity on the Y axis. You can quantify how quickly the ROC curve rises to the upper left hand corner by measuring the area under the curve. The larger the area, the better the diagnostic test is. If the area is 1.0, you have an ideal test because it achieves both 100% sensitivity and 100% specificity. If the area is 0.5, then you have a test that has effectively 50% sensitivity and 50% specificity. This is a test that is no better than flipping a coin. In practice, a diagnostic test is going to have an area somewhere between these two extremes. The closer the area is to 1.0, the better the test is, and the closer the area is to 0.5, the worse the test is.

CHAPTER 4

EXPERIMENTAL RESULTS

In this chapter, the experimental results obtained using the proposed feature selection algorithm are shown, and the performance of the algorithm is assessed by comparing with other algorithms.

4.1 Results of the Proposed Algorithm

Baseline adjustment is done by using a MATLAB function, which is a baseline signal that has to be subtracted and is generated because sometimes the detector overestimates the number of ions arriving at its surface. Smoothing reduces the noise in the spectra and peak detection picks the peaks with intensity values of 10 or more. After these steps the total number of peaks are reduced to a great extent, thereby decreasing the final number of features to be considered in the next step. The biomarker needed to be selected from the preprocessed mass spectrum data i.e., from the back projected individual spectra we should be able to distinguish between cancer spots and normal spots.

From the above preprocessing steps, 820 peaks are obtained for each spectrum. There are a total of 974 spectra, out of which 27 belong to cancer spots, and the rest of them are normal. The proposed feature selection algorithm is applied to the preprocessed prostate cancer data to select the most discriminating peaks. Note that before applying the feature selection algorithm, Fisher score [41] is used to preselect 30 out of the 820 peaks.

The Fisher score is intended to be a measure of the difference between distributions of a single variable. A particular feature's Fisher score is computed by the following formula:

$$F(i) = \frac{\left(\mu_i^+ - \mu_i^-\right)^2}{\left(\sigma_i^+\right)^2 + \left(\sigma_i^-\right)^2} \tag{4-1}$$

where $\mu_i^{\pm}$ is the mean value for the i^{th} feature in the positive or negative profiles, and $\sigma_i^{\pm}$ is the standard deviation. Features with high Fisher scores possess the desirable quality of having a large difference between means of case versus control groups, while maintaining low overall variability. These features are more likely to be consistently expressed differently between case and control samples and therefore indicate good candidates for feature selection.

Table 4.1 shows the sensitivity and specificity obtained by using the selected peaks by our algorithm. This is the training result. The indices for peaks shown in Table 4.1 are ranks based on the Fisher score.

Table 4.1: Feature selection results of the proposed algorithm.

Predicted Acc %	Sensitivity %	Specificity %	Subset Size	Subset Members
99.383984	85.185185	99.788807	1	{1,}
99.486653	88.888889	99.788807	2	{1,7,}
99.794661	100.000000	99.788807	3	{1,6,4,}
99.897331	100.000000	99.894403	4	{1,6,4,29,}
99.897331	100.000000	99.894403	5	{1,6,4,29,8,}
99.691992	92.592593	99.894403	6	{1,6,4,29,8,5,}
99.794661	92.592593	100.000000	7	{1,6,4,29,8,5,3,}

Most of the selected peaks correspond to the *m/z* values around 4000. As seen from Table 4.1, selection of 3 or 4 peaks is the best combination for the prediction task as they

achieved high sensitivity and specificity, so one of the combinations can be used as a biomarker. Inclusion of more peaks is not helpful based on Table 4.1.

To find the exact combination from Table 4.1, a graph is plotted between the number of features included in the classifier and the sensitivity of 5-fold cross validation and is shown in Figure 4.1. Sensitivity is only used for final combination selection because the proposed feature selection algorithm achieved almost 100% specificity for all the combinations.

By the result shown in Figure 4.1 the combination of 3 features is used since it achieved the highest sensitivity. The zoom-in result of Figure 4.1 is shown in Figure 4.2. To show the effectiveness of the feature selection algorithm the distribution of two proteins that are selected by the proposed algorithm are plotted in Figure 4.3 and Figure 4.4. The effectiveness of the algorithm is also shown by visualizing the cancer tissue by using BioMap software. For comparison, the original tissue is shown in Figure 4.5. The cancer affected region in Figure 4.5 is shown in a black circle, and its zoom-in is shown in Figure 4.6. The distribution of the two proteins is also shown in Figure 4.7 and Figure 4.8 using BioMap software. The dark orange color suggests that the detected proteins are present in large quantities in the cancer affected region.

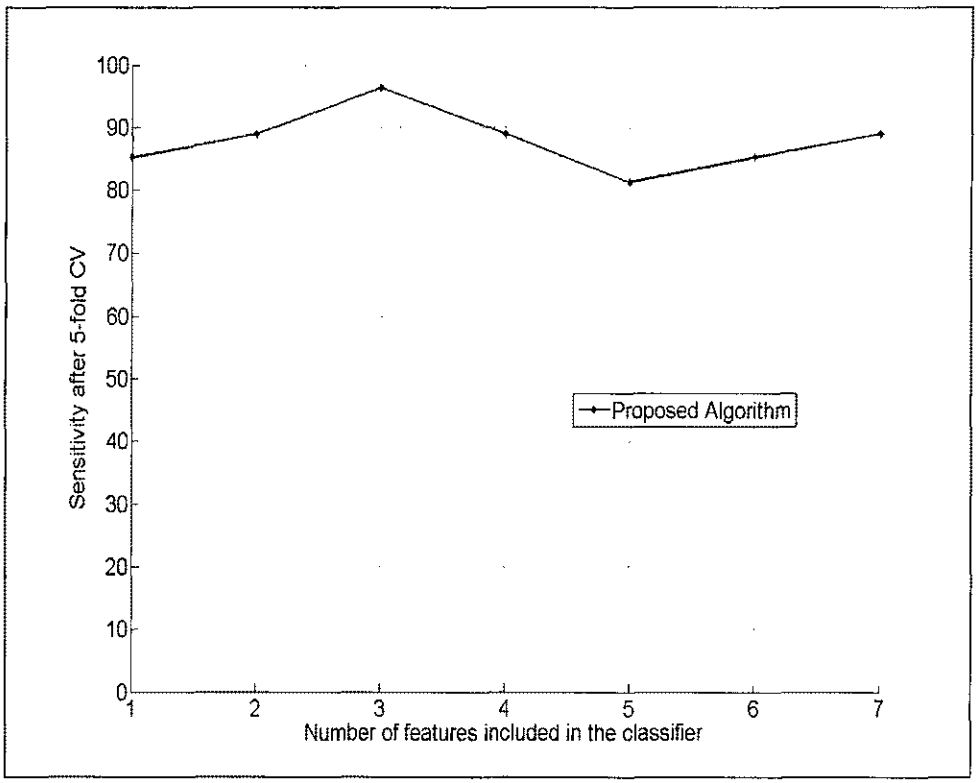

Figure 4.1: Effect of feature numbers.

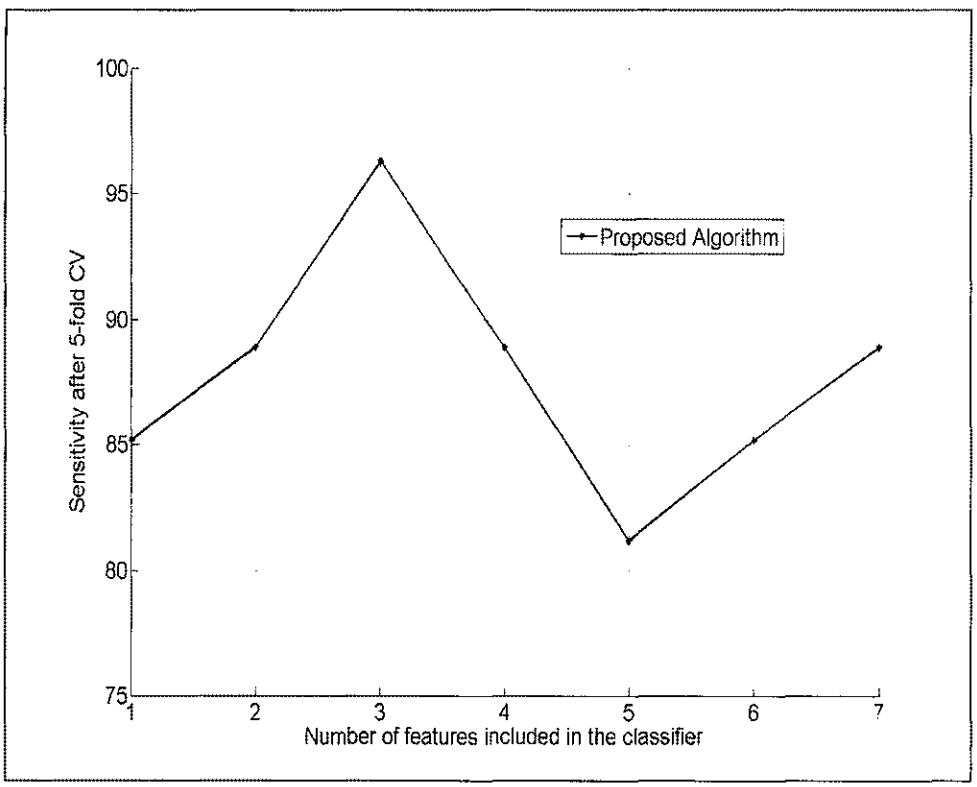

Figure 4.2: Zoom-in of Figure 4.1.

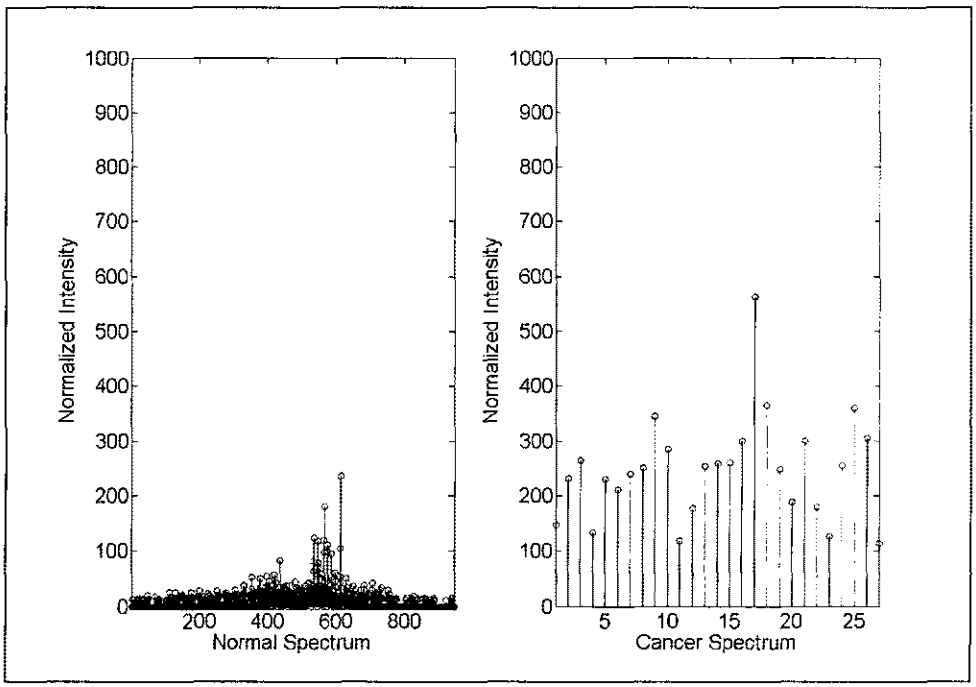

Figure 4.3: Distribution of protein with *m/z* value around 4000 selected by the proposed algorithm.

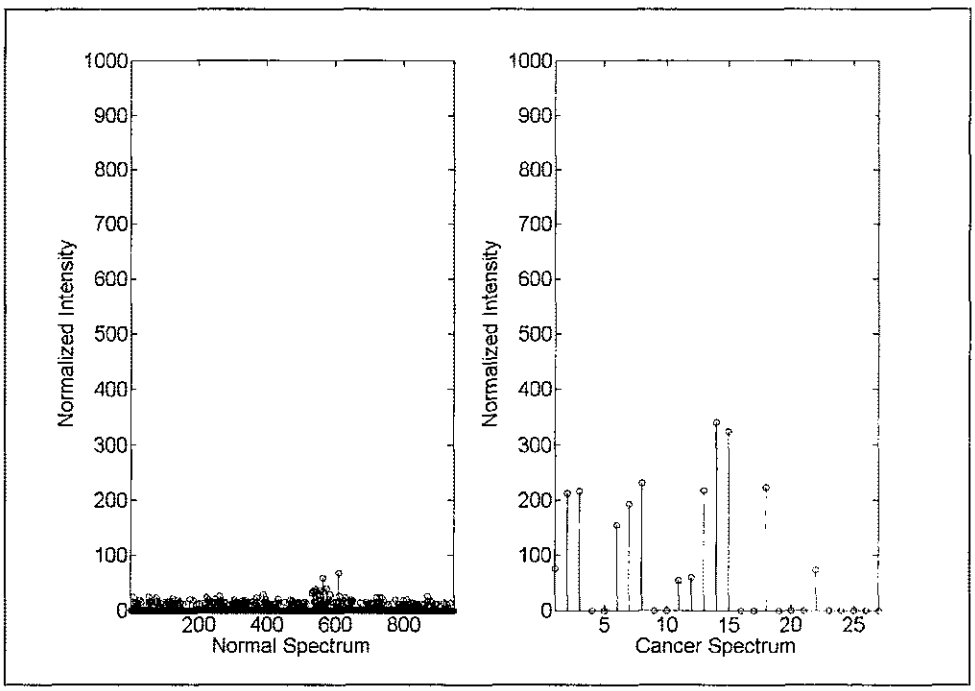

Figure 4.4: Distribution of another protein with *m/z* value around 4000 selected by the proposed algorithm.

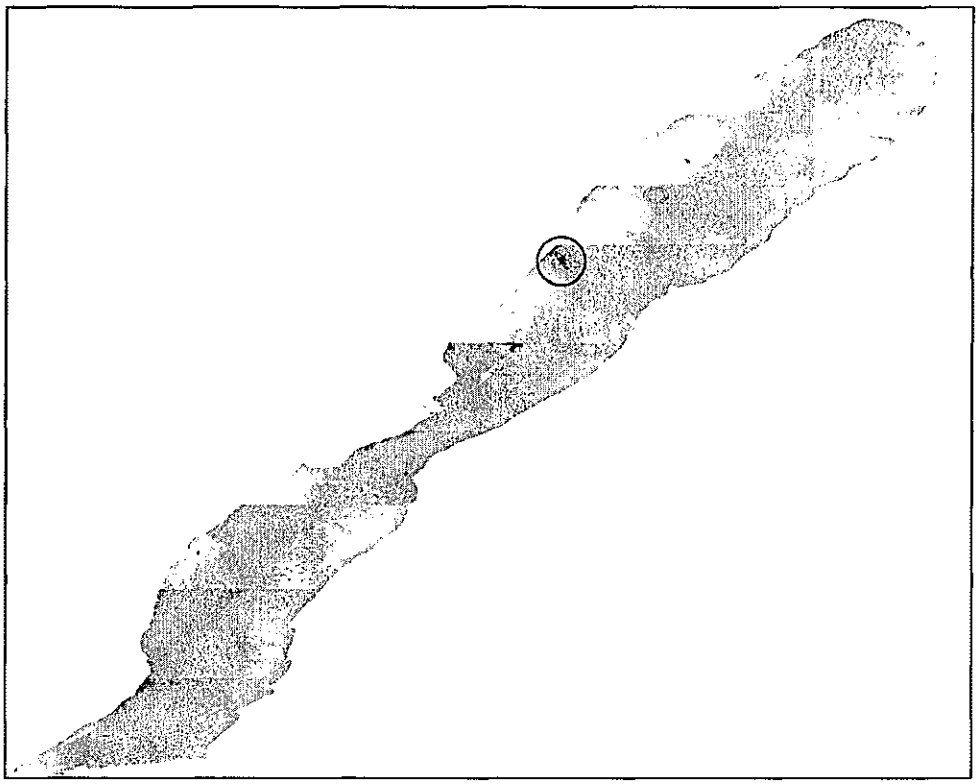

Figure 4.5: Prostate cancer tissue sample.

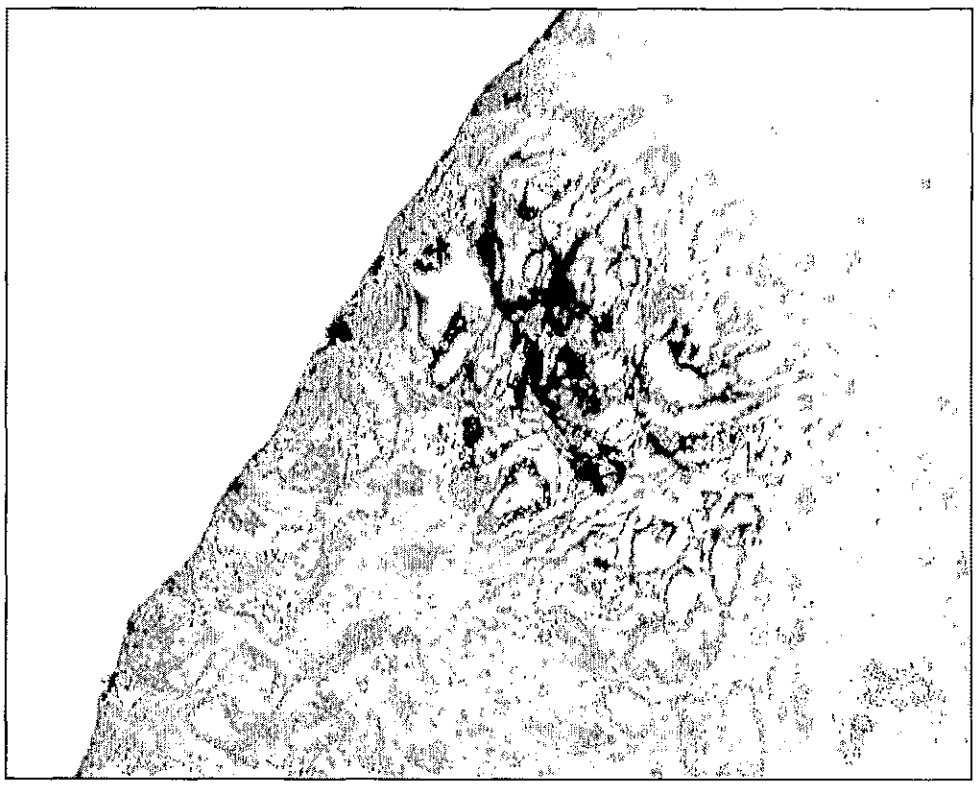

Figure 4.6: Cancer affected area.

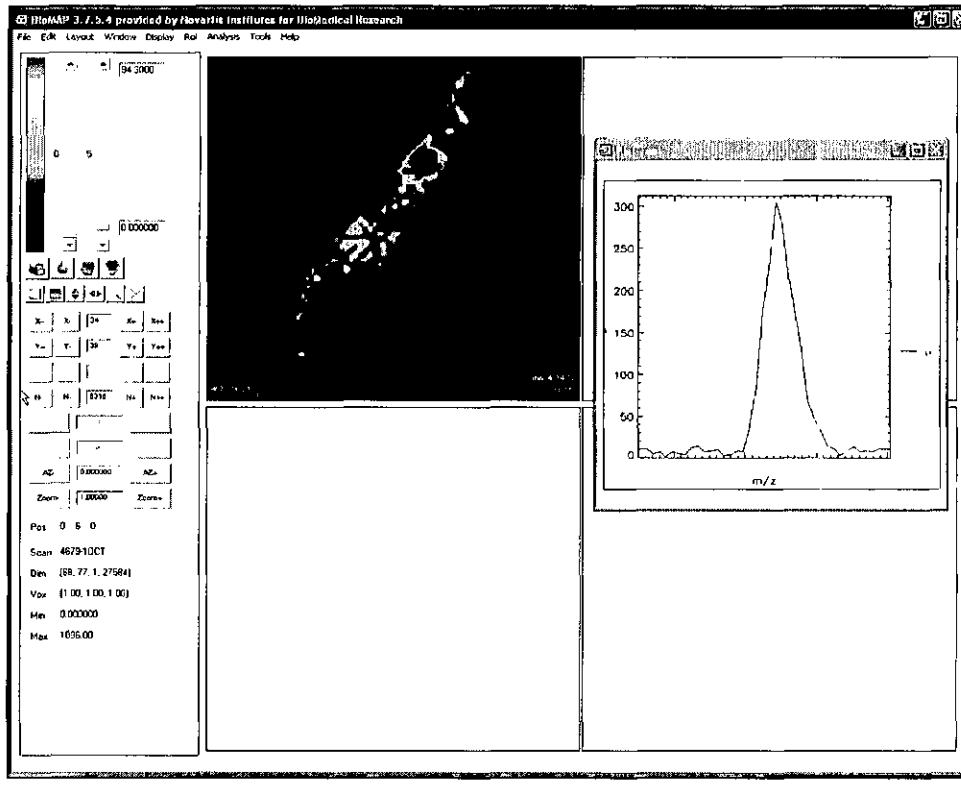

Figure 4.7: Distribution of protein in Figure 4.3 shown by BioMap.

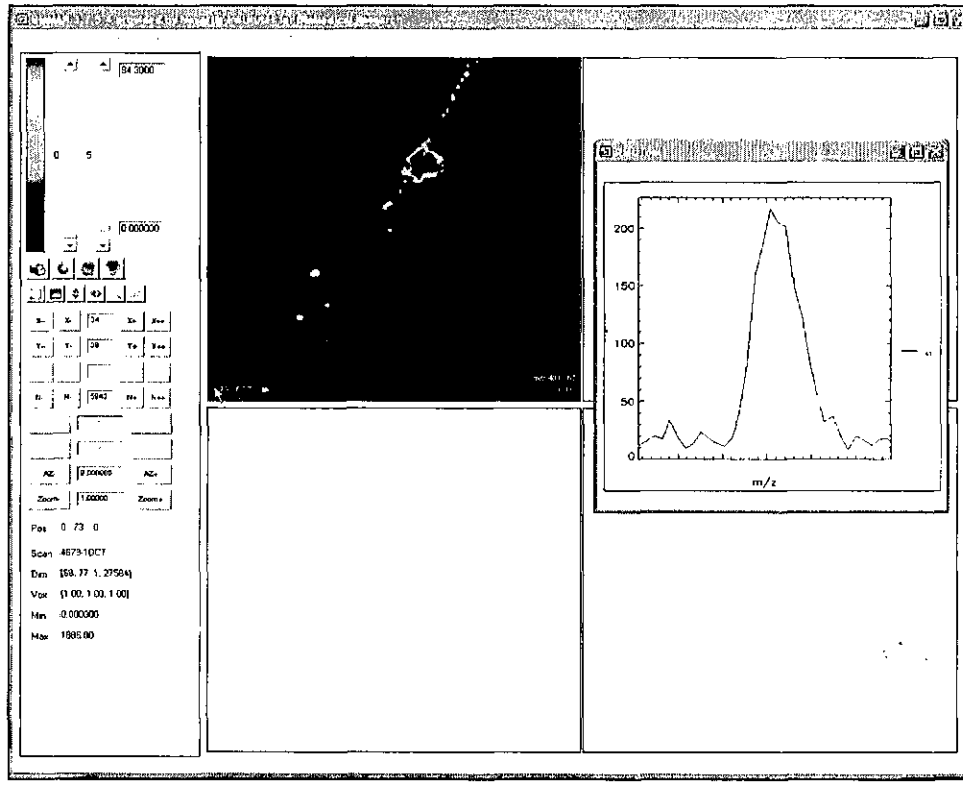

Figure 4.8: Distribution of protein in Figure 4.4 shown by BioMap.

To show the performance of our methodology, the ROC curve is plotted by using 3 features in the classifier and 5-fold CV. This is shown in Figure 4.9. The top plot is the original, and the bottom plot is its zoom-in. The area under the curve close to 1.0 suggests that the proposed feature selection algorithm performed well in identifying the discriminating pattern.

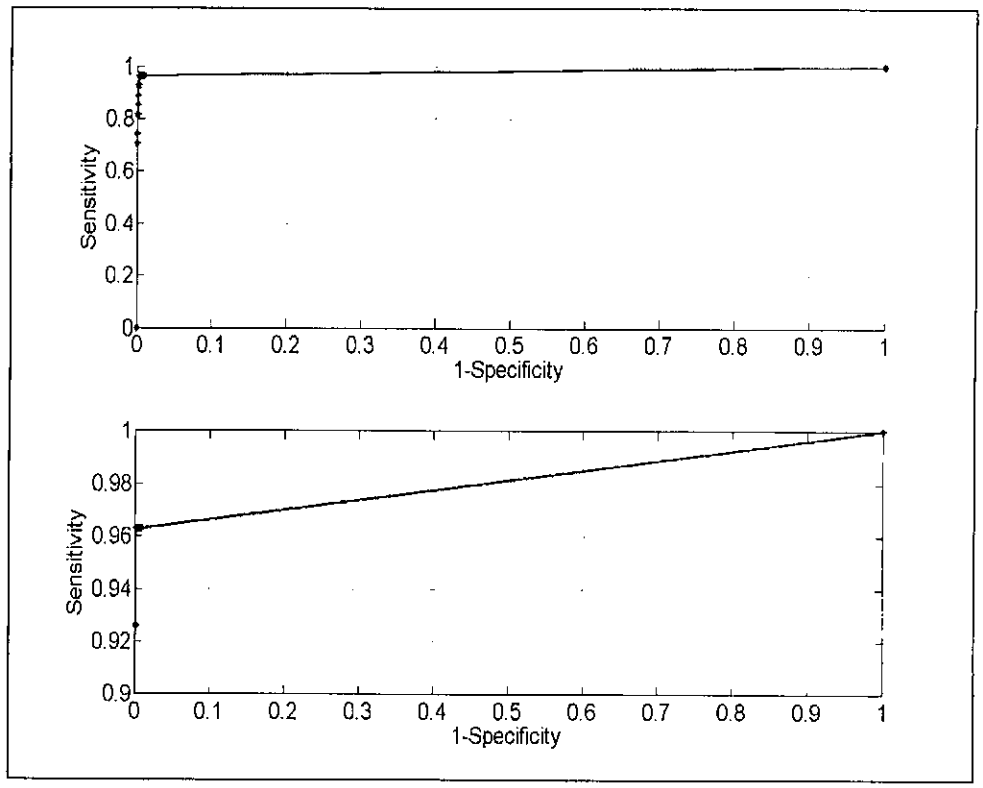

Figure 4.9: ROC curve using 3 peaks.

4.2 Comparison with Other Algorithms

For the purpose of comparison, other famous feature selection algorithms like AUC score, J5test, minimum-Redundancy-Maximum-Relevance, Random Search and Genetic Algorithm are implemented. After the data is preprocessed these algorithms are applied to it to select the compact set of features.

Figure 4.10 shows that the proposed feature selection algorithm is able to detect the minimum number of features with maximum sensitivity when compared to other algorithms implemented. The genetic algorithm gave the highest sensitivity if 4 features are selected for the classification, which is equal to that of the proposed algorithm only using 3 features. In data modeling, we always prefer to use fewer features if similar results can be achieved. There are some redundant features that cannot be helped in improving the classification that were selected by the genetic algorithm. The second advantage of the proposed algorithm is that it can select different combinations in a single run while the genetic algorithm selects only one combination per run, i.e., in order to produce the best combination of 3 features and 4 features, we need to run the genetic algorithm twice. Hence, the proposed algorithm is the best for the identification of the biomarker for the prostate dataset in terms of compactness and computational efficiency.

ROC plots for all the other algorithms are plotted based on the top 3 features selected by the respective algorithms using 5-fold cross validation and are shown in Figure 4.11. The zoom-in is shown in Figure 4.12. These plots show that the proposed algorithm is best not only in selecting the minimum number of features but also in discriminating between the cancer samples from normal samples with high sensitivity and specificity as the area under the ROC curve is highest for the proposed algorithm. Table 4.2 shows the computational times of all the feature selection algorithms in selecting different combinations of features. The table proves the computational efficiency of the proposed algorithm.

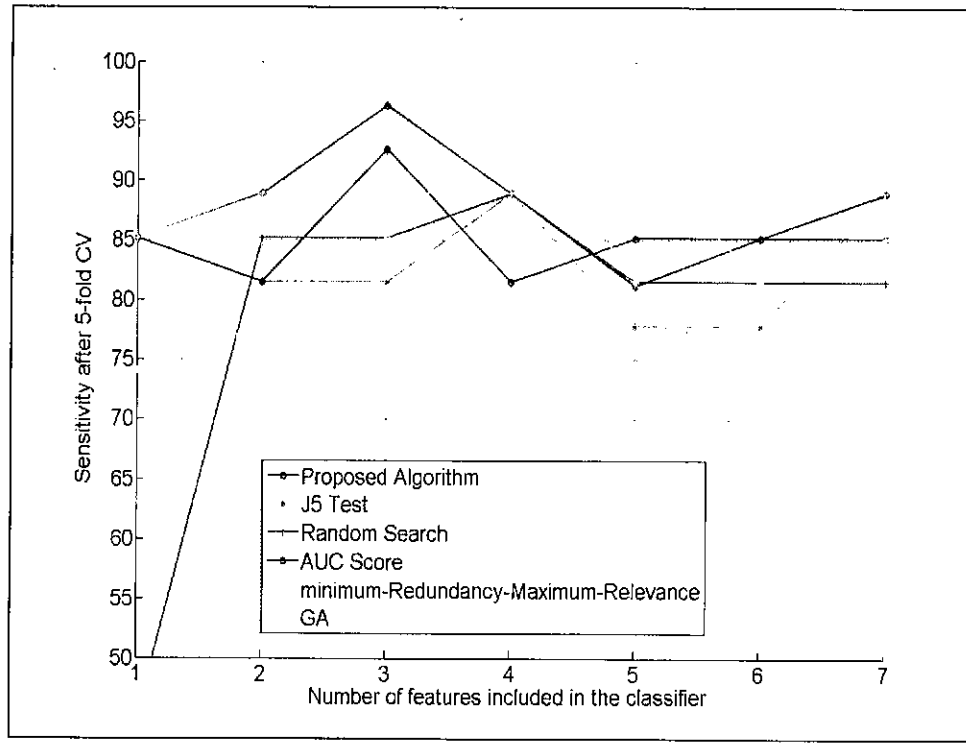

Figure 4.10: Effect of increasing peaks for all the algorithms.

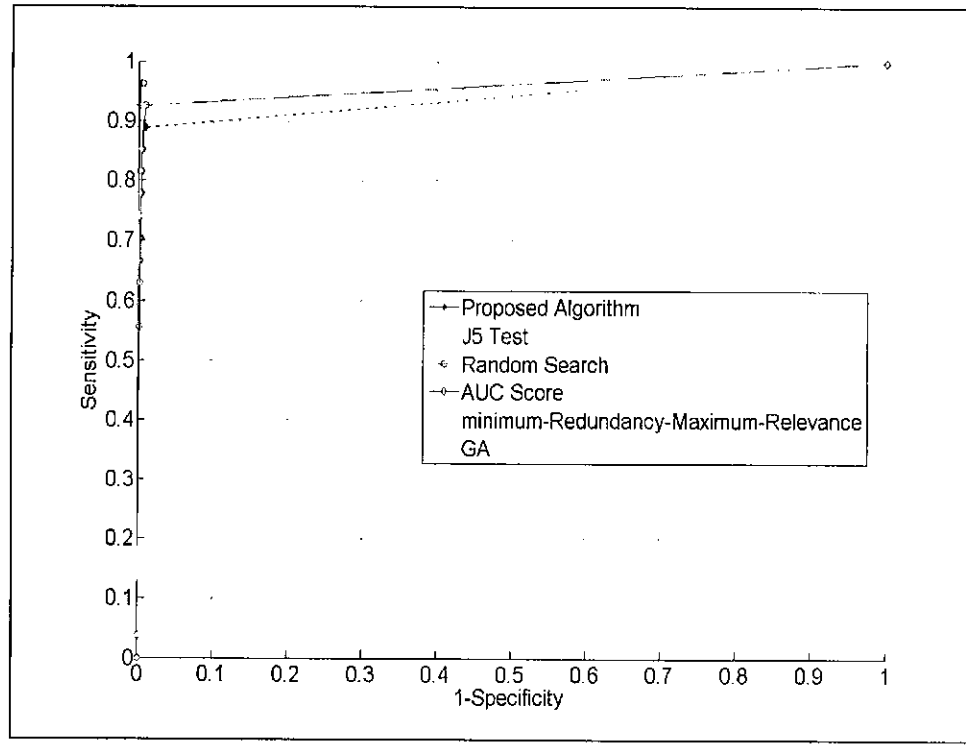

Figure 4.11: ROC plots for all the algorithms.

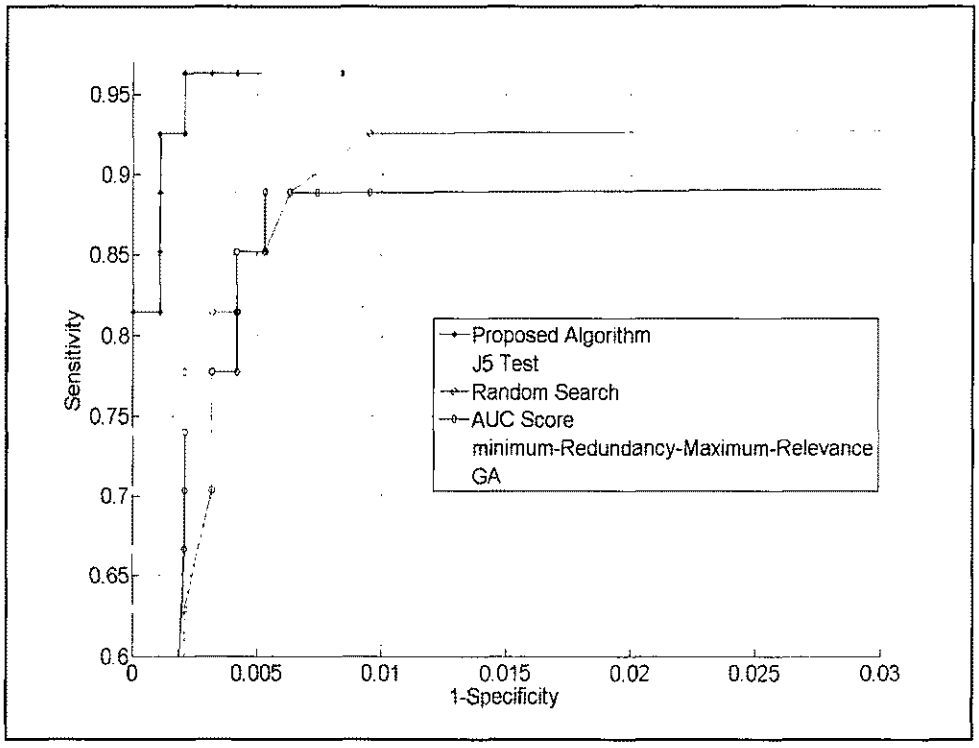

Figure 4.12: Zoom-in of Figure 4.11.

Table 4.2: Time of execution of all the feature selection algorithms.

Algorithm	Feature combinations selected	Computation time in seconds
Proposed algorithm	30	2.96
Genetic algorithm	7	45.09
Random search	30	11.28
mRMR	30	9.93

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

This thesis proposed a three-step pipeline for biomarker identification: (1) Data Preprocessing, (2) Feature selection and (3) Classification. Data preprocessing techniques were applied to the raw mass spectrometric prostate cancer data. Routine functions provided by MATLAB for preprocessing, including baseline correction, noise removal by wavelet methods, peak normalization, peak alignment and peak detection were utilized. The proposed feature selection algorithm was then applied to the preprocessed data for peak selection.

The proposed method is efficient because only one data pass is sufficient for the peak selection. It evaluates peak combinations by considering their interactions, and the correlated peaks will be eliminated automatically. A special procedure, the OR algorithm, was used for unbalanced data handling. The selected compact set of highly discriminative protein peaks was then used by an MLP classifier to classify MS spectra as cancer or normal.

After the preprocessing techniques, we reduced the dimension of raw data from 82,756 to 820. The feature selection algorithm reduced the dimensionality further from 820 to 3. Five-fold cross validation results showed that the developed pipeline achieved a sensitivity of 96.29% and specificity of 99.68% using three peaks selected by the proposed method. The proposed method outperformed many other currently used feature selection algorithms for the identification of the prostate cancer biomarker.

5.2 Future Work

Future work includes validating the proposed method using more MALDI-MSI data sets. Statistical analysis will also be performed.

REFERENCES

[1] Pisani Paola, Bray Freddie, and Parkin 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.

[2] Markus 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.

[3] M. C. Beduschi, and J. E. Oesterling, "Percent free prostate-specific antigen: the next frontier in prostate-specific antigen testing," *Urology*, Vol. 51, pp. 98–109, 1998.

[4] 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.

[5] I. M. Thompson, D. K. Pauler, P. J. Goodman, C. M. Tangen, M. S. Lucia, H. L. Parnes, L. M. Minasian, L. G. Ford, S. M. Lippman, E. D. Crawford, J. J. Crowley, and C. A. Coltman Jr, "Prevalence of prostate cancer among men with a prostate-specific antigen level ≤4.0 ng per milliliter," *New England Journal of Medicine*, Vol. 350, pp. 2239–2246, 2004.

[6] Lisa H. Cazares, *et al.*, "Normal, Benign, Preneoplastic, and Malignant Prostate Cells Have Distinct Protein Expression Profiles Resolved by Surface Enhanced Laser Desorption/Ionization Mass Spectrometry," *Clinical Cancer Research*, Vol. 8, pp. 2541–2552, August 2002.

[7] Xutao Deng, Huimin Geng and Hesham 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] R. R. Drake, L. Cazares, and O. J. Semmes, "Mining the low molecular weight proteome of blood," *Proteomics Clinical Applications*, Vol. 1, pp. 758–768, 2007.

[9] Rowan E. Moore, Jennifer Kirwan, Mary K. Doherty and Phillip D. Whitfield, "Biomarker Discovery in Animal Health and Disease: The Application of Post-Genomic Technologies," *Biomarker Insights*, Vol. 2, pp. 185–196, 2007.

[10] Deukwoo Kwon, Mahlet G. Tadesse, Naijun Sha, Ruth M. Pfeiffer, and Marina Vannucci, "Identifying Biomarkers from Mass Spectrometry Data with Ordinal Outcome," *Cancer Informatics*, Vol. 3, pp. 19–28, 2007.

[11] Edwin M. Posadas, Ben Davidson, and Elise C. Kohn, "Proteomics and ovarian cancer: implications for diagnosis and treatment: a critical review of the recent literature," *Current Opinion in Oncology,* Vol. 16, No. 5, pp. 478-484, September 2004.

[12] J. Lyons-Weilera, R. Pelikanb, H. J. Z. III, D. C. Whitcomb, D. E. Malehorn, W. L. Bigbee, and M. Hauskrecht, "Assessing the statistical significance of the achieved classification error of classifiers constructed using serum peptide profiles, and a prescription for random sampling repeated studies for massive high-throughput genomic and proteomic studies," *Cancer Informatics*, Vol. 1, no. 1, pp. 53–77, 2005.

[13] Christine Laronga, and Richard R. Drake, "Proteomic Approach to Breast Cancer," *Cancer Control*, Vol. 14, No. 4, pp. 360-368, October 2007.

[14] Biomarkers Definitions Working Group, "Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework," *Clinical Pharmacology and Therapeutics*, Vol. 69, pp. 89–95, 2001.

[15] V. Gewin, "Missing the mark," *Nature*, Vol. 449, No. 18, pp. 770–771, 2007.

[16] Dale McLerran, William E. Grizzle, and O. John Semmes, *et al.*, "Analytical validation of serum proteomic profiling for diagnosis of prostate cancer: Sources of sample bias," *Clinical Chemistry*, Vol. 54, No. 1, pp. 44–52, 2008.

[17] Dale McLerran, William E. Grizzle, Ziding Feng, and O. John Semmes, *et al.*, "SELDI-TOF MS whole serum proteomic profiling with IMAC surface does not reliably detect prostate cancer," *Clinical Chemistry*, Vol. 54, No. 1, pp. 53–60, 2008.

[18] O. John Semmes, Ziding Feng, and Bao-Ling Adam, *et al.*, "Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-offlight mass spectrometry for the detection of prostate cancer: I. assessment of platform reproducibility," *Clinical Chemistry*, Vol. 51, No. 1, pp. 102–112, 2005.

[19] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Transaction on Evolutionary Computation*, Vol. 1, No. 1, pp. 67–82, 1997.

[20] Vamsi K. R. Mantena, Jiang Li and Rick McKenzie, "Biomarker Identification by using Mass Spectrum Data", *Capstone conference*, presentation, VMASC 2008.

[21] N. Dossat, A. Mang, J. Solassol, W. Jacot, Ludovic Lhermitte, T. Maudelonde, J. P. Daurs, and N. Molinari, "Comparison of supervised classification methods for protein profiling in cancer diagnosis," *Cancer Informatics*, Vol. 3, pp. 295–305, 2007.

[22] Y. Qu, B. ling Adam, M. Thornquist, J. D. Potter, M. L. Thompson, Y. Yasui, J. Davis, P. F. Schellhammer, M. C. Lisa Cazares, G. L. W. Jr, and Z. Feng, "Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data," *Biometrics*, Vol. 59, No. 1, pp. 143–151, 2003.

[23] M. Raymer, W. Punch, E. Goodman, L. Kuhn., and A. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Transactions on Evolutionary Computation*, Vol. 4, pp. 164–171, 2000.

[24] Steven R. Myers and Md. Yeakub Ali, "Determination of Tobacco Specific Hemoglobin Adducts in Smoking Mothers and New Born Babies by Mass Spectrometry," *Biomarker Insights* Vol. 2, pp. 269-282, 2007.

[25] Bao-Ling Adam, Yinsheng Qu, John Davis, Michael D Ward, Mary Ann Clements, Lisa H Cazares, O John Semmes, Paul F Schelhammer, Yutaka Yasui, Ziding Feng and George L Wright,Jr, "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.

[26] J. Li, M. T. Manry, P. L. Narasimha, and C. Yu, "Feature selection using a piecewise linear network," *IEEE Transaction on Neural Network*, Vol. 17, No. 5, pp. 1101–1105, 2006.

[27] 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.

[28] 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.

[29] S. Patel, J. Lyons-Weiler, "caGEDA a web application for the integrated analysis of global expression patterns in cancer," *Application of Bioinformatics*, Vol. 3, No. 1, pp. 49-62, 2005.

[30] Hanchuan Peng, Fuhui Long, and Chris 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, August 2005.

[31] L. Li, *et al.*, "Application of the GA/KNN method to SELDI proteomics data", *Bioinformatics*, Vol. 20, No. 10, pp. 1638-1640, 2004.

[32] Steven R. Myers and Md. Yeakub Ali, "Determination of Tobacco Specific Hemoglobin Adducts in Smoking Mothers and New Born Babies by Mass Spectrometry," *Biomarker Insights*, Vol. 2, pp. 269–282, 2007.

[33] W. M. Old, K. Meyer-Arendt, L. Aveline-Wolf, K. G. Pierce, A. Mendoza, J. R. Sevinsky, K. A. Resing, and N. G. Ahn, "Comparison of label-free methods for quantifying human proteins by shotgun proteomics," *Molecular Cellular Proteomics*, Vol. 4, No. 10, pp. 1487-502, 2005.

[34] Kate W. Jordan, Wenlei He, Elkan F. Halpern, Chin-Lee Wu and Leo 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.

[35] 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, August 2007.

[36] Jiang Li, Michael T. Manry, Li-Min Liu, Changhua Yu, and John Wei, "Iterative improvement of neural classifiers," *Proceedings of the Seventeenth International Conference of the Florida AI Research Society*, pp. 700-705, May 2004.

[37] R. G. Gore, Jiang Li, M. T. Manry, L. M. Liu, and Changhua 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.

[38] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, pp. 311-325, March 2006.

[39] 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* Vol. 1, pp. 1034-1038, 2002.

[40] G. D. Magoulas, M. N. Vrahatis, and G. S. Androulakis, "Improving the convergence of the backpropagation algorithm using learning adaptation methods," *Neural Computation*, Vol. 11, pp. 1769-1796, 1999.

[41] C. Bishop, *Neural networks for pattern recognition.* New York: Oxford University Press, pp. 105-112, 1995.

VAMSI KRISHNAM RAJU MANTENA
Department of Electrical and Computer Engineering
Old Dominion University
757-275-5624
vmant002@odu.edu

Education:

Master of Science in **Electrical Engineering**

Old Dominion University, Norfolk, VA. (December 2008)

GPA: **3.85/4.0**

Bachelor of Engineering in **Electronics and Communication Engineering**

Andhra University, India. (April 2006)

GPA: **3.6/4.0**

Research Publications:

- S. Jakkula, V. Mantena, R. Pedada, Y. Shen, and J. Li, "Seasonal Adaptation of Vegetation Color in Satellite Images," *2008 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, (Las Vegas, NV), July 2008.
- V. Mantena, R. Pedada, S. Jakkula, Y. Shen, and J. Li, "Vegetation Identification Based on Satellite Imagery," *2008 International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV)*, (Las Vegas, NV), July 2008.
- Vamsi K. R. Mantena, Jiang Li and Rick McKenzie, "Biomarker Identification by using Mass Spectrum Data", *Capstone conference*, VMASC 2008.
- Vamsi K. R. Mantena, Jiang Li, Rick McKenzie, Lisa Cazares, Richard Drake and John Semmes, "An Efficient Algorithm for Biomarker Identification", *(poster), EDRN workshop*, Bethesda, Maryland 2008.
- Vamsi Mantena, Jiang Li and Yuzhong Shen, "Feature Classification in Aerial and Satellite Imagery Using Clustering Algorithm and Morphological Operations," *ODU-NSU-EVMS-VTC Research Exposition Day: Research Expo*, Norfolk, VA, (Poster), 2008.