Chapter 8
A New Wavelet-Based Approach for Mass Spectrometry Data Classification

Achraf Cohen, Chaimaa Messaoudi, and Hassan Badir

8.1 Introduction

Application of new technologies of big data and statistical learning theory to mass spectrometry data classification problem can have a valuable impact on public health. This need is particularly critical in early detection and identification of cancer. Many strategies can be implemented to combat cancer such as early detection, close monitoring of the patient after initial treatment, and others (Diamandis 2004). Proteomic patterns through mass spectrometry techniques have shown a promising strategy to diagnose cancer.

Mass Spectrometry (MS) is an analytical chemistry technique that was introduced to help to identify the amount and type of chemicals present in a sample by measuring the mass-to-charge ratio and abundance of gas-phase ions. The mass spectrometers consist of three principal elements: an ion source, a mass analyzer, and an ion detection system (Aebersold and Mann 2003). The ionization is the first step in mass spectrometry analysis. The second step is the separation of the ions according to their mass to charge ratio. Finally, the compounds are detected and the relative abundance of each of the resolved ionic species is recorded. The output of the detector is a mass spectrum presented in a plot of the relative abundance or relative intensity as a function of the mass-to-charge ratio, see Fig. 8.1.

A. Cohen (✉)
Department of Mathematics and Statistics, University of West Florida, Pensacola, FL, USA
e-mail: acohen@uwf.edu

C. Messaoudi · H. Badir
National School of Applied Sciences-Tangier, ENSAT, Abdelmalek Essaadi University, Tangier, Morocco
e-mail: messaoudi@ensat.ac.ma; badir.hassan@uae.ma

Y. Zhao, D.-G. Chen (eds.), *New Frontiers of Biostatistics and Bioinformatics*,
ICSA Book Series in Statistics, https://doi.org/10.1007/978-3-319-99389-8_8

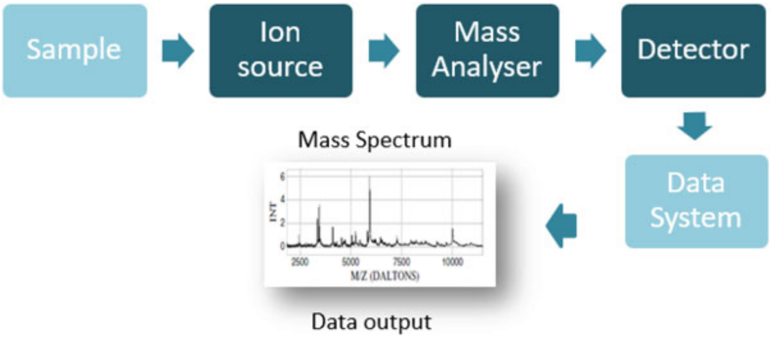

Fig. 8.1 Components of a mass spectrometer

In the last decade, MS data analysis has become an increasingly prominent field allowing the identification, quantification, and characterization of peptides and proteins in biological samples. It has been applied to discover patterns of differentially expressed protein in clinical samples such as blood serum. Especially, biomarker identification that can be used for diagnosis and monitoring of many diseases (Cravatt et al. 2007). The Matrix Assisted Laser Desorption/Ionization Time-Of-Flight (MALDI-TOF) and Surfaced-Enhanced Laser Desorption/Ionization Time-Of-Flight (SELDI-TOF) are high-throughput technologies for the acquisition of protein expression profiles from biological fluids (serum, plasma, etc.). The use of these technologies, with statistical modeling, is essential for (1) the identification of novel protein biomarkers of disease and (2) the classification of a new unseen mass spectrum.

MS data are given by the number of mass-to-charge ratios (m/z). Tens of thousands of m/z are available in the data but not necessarily all are used to MS data classification. It is reasonable to have a feature extraction procedure that is able to decrease the effects of noise, reduce dimension, and define new features to represent the data. These new features are used to develop a good classification model (Das 2001). In the last decades, MS data analysis for cancer identification has focused on two main concepts (1) selecting features from MS spectrum and (2) developing classification models for prediction. Both concepts should work together in order to provide an accurate model for classifying MS data.

The conventional method for processing an MS spectrum is to perform a number of preprocessing steps before developing any statistical models. These tools include baseline correction, normalization, and denoising (Dubitzky et al. 2007, pp. 79–102). The authors in Petricoin et al. (2002) developed a bioinformatics tool to identify ovarian cancer using self-organizing clustering analysis and genetic algorithm. In Tang et al. (2010), the authors proposed an approach for dimensionality reduction and tested it using mass spectrometry data for ovarian cancer. They used the mean, variance, skewness, and kurtosis in order to reduce the dimension. A Kernel Partial Least Squares model is then developed for ovarian cancer classification. Moreover, Li and Zeng (2016) proposed a method based on the model of uncorrelated linear discriminant analysis combined with variable selection method, applied to serum SELDI-TOF MS for ovarian cancer identification. Wu et al. (2016) proposed a classification model based on probabilistic principal component analysis and support vector machine. The model was applied to ovarian cancer. Sharma and Singh (2016) suggested the use of the neural network for diagnosis of ovarian cancer.

de Noo et al. (2006) studied colorectal cancer using the MALDI-TOF serum. In a randomized block design, pre-operative samples from 66 colorectal cancer patients and 50 controls were used, and a classification model is built using a linear discriminant analysis with double cross-validation. Another study on colorectal cancer is given in Ward et al. (2008). The authors used a logistic regression model to classify MS spectrum for 67 patients with colorectal cancer and 72 non-cancer control subjects. Lung cancer was studied in Yildiz et al. (2007), the paper investigated MS data to identify lung cancer cases from matched controls. MALDI-MS data were used with two methods of analysis: the weighted flexible compound covariate method and support vector machine.

Pancreatic cancer was the goal of the study given in Ge and Wong (2008). The authors investigated the utility of three feature selection schemas Student t-test, Wilcoxon rank sum test, and genetic algorithm. Some of the selected features were then used to classify MS Pancreatic cancer through six different decision tree classifier ensembles, such as Random forest, Adaboost, and others. Ohn et al. (2016) used 2D polyacrylamide gel electrophoresis (2D PAGE) approach to generate the 2D proteome patterns, and they then compared three classification methods: genetic algorithm combined with SVM, stepwise forward feature selection with K-NN, and random forest. These methods were applied to identify breast cancer.

Lancashire et al. (2009) presented a review of the concepts related to neural networks with their applications in mass spectrometry and focus on cancer studies. In this study (Gromski et al. 2014), the researchers compared feature selection methods with some classification approaches such as Random Forest with its variable selection techniques and SVM combined with support vector machines-recursive feature elimination, and they showed better performance is given by SVM. Awedat et al. (2016) proposed a compressive sensing sampling approach to reducing the dimension. They showed L2-algorithm with regularization terms has better performance than standalone L2-algorithm.

Wavelet analysis has been shown potential application for MS classification to (a) reducing dimension, (b) extracting features, or (c) denoising data. In Yu et al. (2005), a procedure to classify ovarian cancer based on MS data was developed. The authors combined binning, Kolmogorov-Smirnov test, wavelet analysis, and support vector machines to preprocess and develop a classification model, The authors used the *db4* wavelet. Another classification approach of proteomic MS data based on bi-orthogonal discrete wavelet transform and support vector machines was proposed in Schleif et al. (2009). The authors used *bior3.7* wavelet for denoising purposes. Du et al. (2009) proposed a workflow for MS classification based on wavelet analysis, Kolmogorov-Smirnov test with bagging predictor. The wavelet *sym8* was used to denoise the MS data. Nguyen et al. (2015) showed that combining *Haar* wavelet coefficients and genetic algorithm provides a good selection feature subset for the performance classification, but genetic algorithms require a random initialization that may lead to different results. The Wavelet-based function mixed model, which generalizes the linear mixed models to the case of functional data was used in order to analyze MS-data (Morris et al. 2006).

The goal of this chapter is to present a new approach for MS data classification. The proposed approach is original and based on a combination between principal component and wavelet analyses in addition to a new T^2 statistic. Most of the previous research using wavelet analysis did not show how they did select the wavelet family for their analysis. To this end we propose a prior study to select the best-suited wavelet for the analysis. This will help future MS research to have a subjective tool for wavelet selection. The principal component analysis is applied to six features (statistics) that are calculated on the wavelets coefficients (approximation and details). Next, we propose a new statistic $T^2 = \sqrt{T_a^2 + T_d^2}$ combining T^2 on the approximation and details coefficients, respectively. Finally, a support vector machine model is built on the new aforementioned statistic. The proposed approach shows high accuracy, specificity, and sensitivity. We provide a detailed description of each step to ensure the reproductivity of the present research work.

This chapter is organized as follows. Section 8.2 presents the proposed approach for mass spectrometry data classification. In Sect. 8.3, experiments and results are given, and Sect. 8.4 presents conclusions and some directions of research.

8.2 The Proposed Approach

We have designed and implemented a new approach to classify mass spectrometry data. The main steps of our proposed approach are illustrated in Fig. 8.2. The philosophy of the method consists of subdividing the MS sample into several windows and extracting from them some features that will help discriminate/classify the entire MS spectrum. The wavelets analysis has potential capabilities to extract features especially when noisy data is used such as the case with MS data. The principal component analysis with T^2 statistic is used in order to aggregate the features from the wavelets coefficients into one statistic.

The proposed approach can be implemented as follows:

In this approach, each MS spectrum is represented by a T^2 statistic calculated into the feature space of the principal component analysis. The latter is applied to the features (Energy, Mean, Kurtosis, Skewness, Variance, and Coefficient of Variation) of the wavelets coefficients. This approach will be applied to a real dataset in the next section.

1: Input data: MS spectrum X of length L

2: Divide X into n samples of length $N = 2^J$, and rearrange X into a data matrix $Z_{(N \times n)}$

3: Apply Discrete Wavelet Transform using *bior3.1* to each column of $Z_{(N \times n)}$; see Sect. 8.2.1

4: Compute [Energy, Mean, Kurtosis, Skewness, Variance, and Coefficient of Variation (CV)] of the approximation and details wavelets coefficients as follows:

$$Energy = \sum w_i^2; \qquad Mean = \sum w_i/m \tag{8.1}$$

$$Variance = \sum (w_i - Mean)^2/(m-1); \qquad CV = \frac{\sqrt{Variance}}{Mean} \tag{8.2}$$

$$Skewness = E\left[\left(\frac{X-\mu}{\sqrt{Variance}}\right)^3\right]; \qquad Kurtosis = E\left[\left(\frac{X-\mu}{\sqrt{Variance}}\right)^4\right] \tag{8.3}$$

The data look now as follows, for both approximation and details coefficients:

$$\textbf{Feature} = \begin{array}{ccccccc} & \textbf{Energy} & \textbf{Mean} & \textbf{Variance} & \textbf{Skewness} & \textbf{Kurtosis} & \textbf{CV} \\ Window_1 & E_1 & M_1 & V_1 & Sk_1 & Ku_1 & CV_1 \\ Window_2 & E_2 & M_2 & V_2 & Sk_2 & Ku_2 & CV_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Window_n & E_n & M_n & V_n & Sk_n & Ku_n & CV_n \end{array}$$

5: Apply Principal Component Analysis to $Feature_{a_j}$ and $Feature_{d_j}$ data matrices, and select a number of principal components (reduced space)

6: Compute $T^2_{a_j}$ and $T^2_{d_j}$ statistics corresponding to the approximation and details wavelets coefficients, respectively, in the reduced space, see Sect. 8.2.2

7: Develop an SVM model on the $T^2 = \sqrt{T^2_{a_j} + T^2_{d_j}}$ statistic, see Sect. 8.2.3

where w_i are the wavelet coefficients (either approximations or details), m is the number of coefficients, L is the length of the MS, J is a natural number, and a_j and d_j are the approximations and details wavelet coefficients, respectively.

8.2.1 Wavelets Analysis

Wavelet analysis is a mathematical tool that consists of projecting data into a time-frequency representation. The theory of Multi-Resolution Analysis (MRA) has linked wavelets theory to filter analysis. It opened the door to apply wavelets to image processing and also resulted in the implementation of the Fast Wavelet Transform (FWT) algorithm (Mallat 1989; Misiti et al. 1996). Wavelets functions are grouped by families such as Haar, Daubechies, Coiflet, Symlet, and Biorthogonal (Daubechies 1992). The Continuous Wavelet Transform (CWT) is a redundant transformation since the scale and the translation parameters are changed

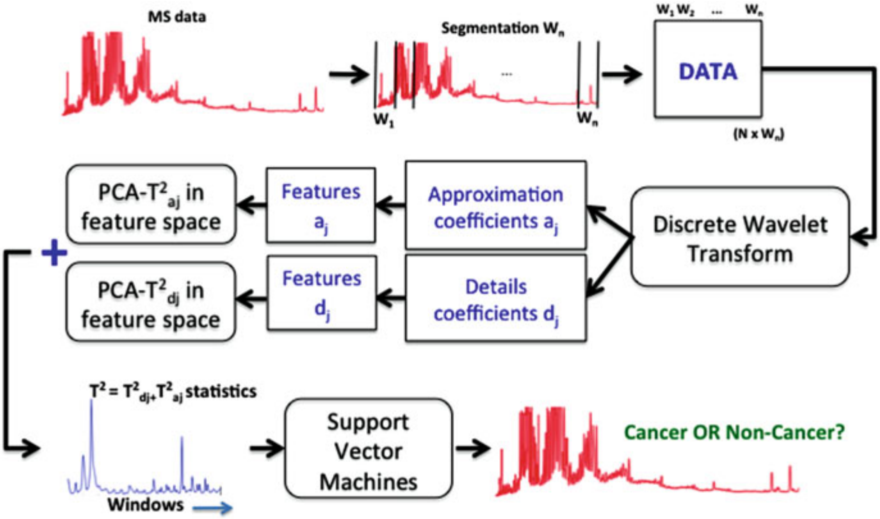
MS data
Segmentation W$_n$
W$_1$
W$_n$
W$_1$ W$_2$ ... W$_n$
DATA
(N x W$_n$)
Discrete Wavelet Transform
Approximation coefficients a$_j$
Details coefficients d$_j$
Features a$_j$
Features d$_j$
PCA-T$^2_{aj}$ in feature space
PCA-T$^2_{dj}$ in feature space
+
T^2 = T$^2_{dj+}$T$^2_{aj}$ statistics
Windows
Support Vector Machines
Cancer OR Non-Cancer?

Fig. 8.2 The proposed method for MS classification

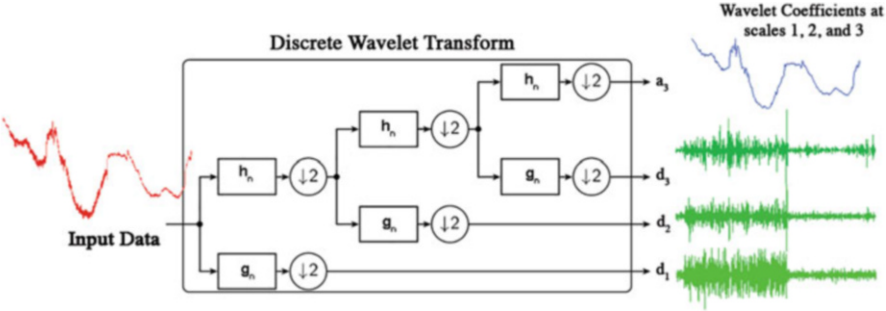
Discrete Wavelet Transform
Wavelet Coefficients at scales 1, 2, and 3
Input Data
h$_n$
g$_n$
↓2
a$_3$
d$_3$
d$_2$
d$_1$

Fig. 8.3 The discrete wavelet transform through filter banks

continuously. The Discrete Wavelet Transform (DWT) is computationally efficient and can be achieved by the discretization of the scale s and translation τ parameters, as follows:

$$\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k) \tag{8.4}$$

where $s = 2^j$ and $\tau = ks$; $j, k \in \mathbf{Z}$.

These wavelet bases are orthogonal and defined in the framework of the Multi-Resolution Analysis (MRA), which provides a multiscale decomposition using orthogonal wavelets families across filter banks, see Fig. 8.3.

The wavelets coefficients of the DWT, approximations $a_j(k)$ and details $d_j(k)$, are given as follows:

$$a_j(k) = \sum_{i=0}^{l} h[i]a_{j-1}[2k-i] \tag{8.5}$$

$$d_j(k) = \sum_{i=0}^{l} g[i]a_{j-1}[2k-i] \tag{8.6}$$

where $a_0 = x$ the original signal, j represents the decomposition scale; $k \in Z$; l is the filter length; h and g are the scaling and wavelets filters, respectively.

The past research publications in bioinformatics have used the shrinkage techniques, which consist of thresholding wavelets coefficients. These techniques have shown a good performance for reducing noise in mass spectrometry data. Several thresholds have been developed, VisuShrink (Donoho and Johnstone 1994), RiskShrink, SUREShrink (Donoho and Johnstone 1995; Donoho 1995), FirmShrink (Gao et al. 1997; Gao 1998), to name a few. One of the benefits of the wavelet transform is the plenty of the wavelets functions developed over the past decades, but from such advantage arises the question of how to select a wavelet that is best suited for analyzing MS data.

There are two approaches in order to choose a wavelet for a specific signal. First, the qualitative methods such as *orthogonality*, *symmetry*, and *compact support*. Second, the quantitative measures such as *energy*, *entropy*, *mutual Information*, *conditional entropy*, and *energy-to-Shannon entropy ratio*. In this work, we used the *energy-to-Shannon entropy ratio*, which is defined as:

$$R = \frac{Energy}{Entropy} = \frac{\sum^N \mid wt(s,i) \mid^2}{-\sum^N p_i \log_2 p_i} \tag{8.7}$$

where N is the number of wavelet coefficients and wt represents the wavelets coefficients, s is the scaling parameter, and $p_i = \frac{|wt(s,i)|^2}{Energy}$.

The set of wavelets that has given a large *energy-to-Shannon entropy ratio* should be considered the candidate wavelets, one can choose the wavelets that have produced the largest *energy-to-Shannon entropy ratio*.

We conducted a preliminary study to choose which wavelet will be used to extract features. We considered 58 wavelets, and by using the Breast cancer Mass Spectrometry data presented in Sect. 8.3. The largest average *energy-to-Shannon entropy ratio* is equal to 49.88 and given by *bior3.1*, see Fig. 8.4. Therefore, we chose the biorthogonal wavelet (Cohen et al. 1992) *bior3.1* as the best-suited wavelet of the analysis.

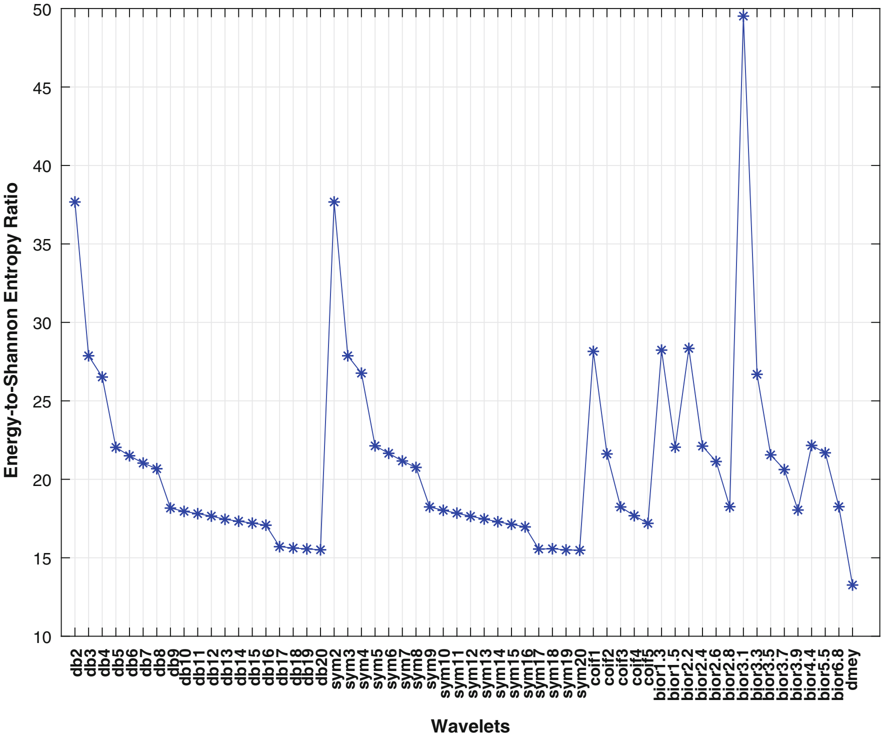

Fig. 8.4 The average energy-to-Shannon entropy ratio using 30 MS Spectrum and 58 Wavelets. dbN: Daubechies of order N; Sym: Symlet; Coif: Coiflet; bior: Biorthogonal, dmey: discrete Meyer

8.2.2 Principal Component Analysis and Hotelling T^2 Statistic

Principal component analysis (PCA) (Jolliffe 1986) is widely used for data exploration and interpretation. Principal component analysis of a data matrix provides new uncorrelated variables (principal components) whose variances are as large as possible. Consider a normalized data matrix, with p variables and N observations.

$$\mathbf{Z} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{Np} \end{pmatrix} \tag{8.8}$$

The covariance matrix of Z can be approximated as:

$$\hat{\Sigma} = \frac{1}{N-1} Z^T Z = P \Lambda P^T \tag{8.9}$$

where $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ with $\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_p$. λ_i are the eigenvalues and P are the eigenvectors of $\hat{\Sigma}$. According to λ_i's, P and Λ could be divided into a feature space (*feat*) and a residual space (*res*). We can then rewrite P and Λ as follows:

$$P = \left[\, P_{feat} \;\; P_{res} \,\right] \tag{8.10}$$

$$\Lambda = \begin{bmatrix} \Lambda_{feat} & 0 \\ 0 & \Lambda_{res} \end{bmatrix} \tag{8.11}$$

The Hotelling T^2 statistic can then be computed as follows:

$$T^2 = Z P_{feat} \Lambda_{feat}^{-1} P_{feat}^T Z^T \tag{8.12}$$

where T^2 is the Hotelling statistic calculated into the multivariate feature space of the principal component analysis, and P^T is the transpose of P. The number of components in the feature space can be determined by using techniques such as the cumulative explained variance and the scree plot. In our approach, the number of principal component in the feature space is determined by the cumulative explained variance technique.

8.2.3 Support Vector Machines

The Support Vector Machines (SVM) are one of the most used statistical learning methods for classification and regression. Classification using SVM can handle problems where a training set *S* is linearly separable or linearly non-separable. The kernel approach allows the training data to be projected into a higher dimensional feature space where the data become separable (Shawe-Taylor and Cristianini 2004; Cristianini and Shawe-Taylor 2000; Vapnik 2013). This property makes the use of SVM valuable for many applications.

Given a training set of pairs (X_i, Y_i), $i = 1, \ldots, N$ where $X_i \in R^n$ and $Y \in \{1(cancer), -1(Non - cancer)\}^N$, the support vector machines find a hyperplane that separates the two classes. The generalized optimal separating hyperplane is determined by w that minimizes the following optimization problem (Cortes and Vapnik 1995):

$$\frac{1}{2} w^T w + C \sum_{i=1}^{N} \epsilon_i \tag{8.13}$$

subject to

$$y_i(w^T \phi(x_i) + b) \geq 1 - \epsilon_i; \qquad \epsilon_i \geq 0 \tag{8.14}$$

where the training set x_i is mapped into a higher dimension space using the function ϕ. C is a positive constant (regularization parameter). It is shown that the problem presented in Eqs. (8.13)–(8.14) depends only on the inner product of x. Therefore, the inner product in the high dimensional feature space can be performed in the input space via the kernel functions. Many kernel functions exist such as polynomial and Gaussian Radial Basis Function (RBF). The Gaussian kernel has given a great attention and defined by:

$$K(x, x') = exp\Bigg(- \gamma \mid\mid x - x' \mid\mid^2 \Bigg) \tag{8.15}$$

where γ is the Kernel parameter. In order to find the best decision boundaries, the hyperparameters C and γ should be controlled. The hyperparameter optimization can be used to select C and the Kernel parameter. Methods such as grid search, random search, and Bayesian optimization are often used for such purpose. In this work, we used a grid search on $C \in \{2^{-8}, 2^{-7}, \ldots, 2^8\}$ and $\gamma \in \{2^{-10}, 2^{-9}, \ldots, 2^{10}\}$ to select the hyperparameters C and γ. We run a 50-fold cross validation on the training data set to determine the hyperparameters as follows: C = 2 and $\gamma = 0.0009765$

8.3 Experiments and Results

The data used are low-mass range SELDI spectra derived from patients with breast cancer and from normal controls. They can be found online at the Department of Bioinformatics and Computational Biology at the University of Texas M.D. Anderson Cancer Center (P. datasets for Breast Cancer 2004). The datasets were generated using IMAC-3 protein chip and sample application was performed using the Biomek 2000 Laboratory Automation Workstation robot (Beckman Coulter, Fullerton, CA). There are 33,885 m/z values and 156 samples where control (normal) patients contribute with 57 samples and 99 samples are cancer. The analysis was done using Matlab and R. The authors are willing to share the code used in this work if requested by e-mail. An example of a sample of cancer and a sample from normal patients is in Fig. 8.5.

In the experiment, each MS sample is subdivided into 64 windows of $2^8 = 256$ observations that is 32,768 m/z values. Since the data have 33,885 values the remaining values are not used. The data then are arranged into a matrix data of 256 rows and 64 columns. Next, the discrete wavelet transform is applied to each column (window) using the *bior3.1* wavelet as shown in Sect. 8.2.1. This results in obtaining the approximation coefficients and details coefficients for each window. Afterwards, the features presented in Sect. 8.2 are computed for each window, therefore the

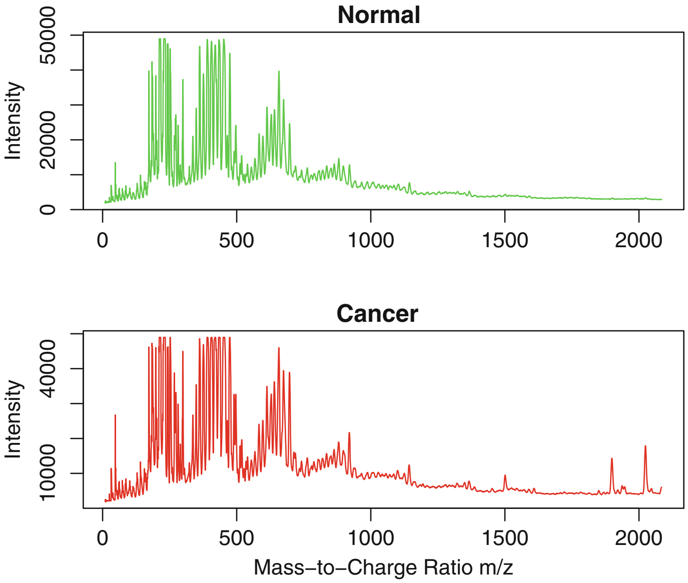

Fig. 8.5 An example of MS samples from cancer and normal patients

data are now arranged as two matrices data of 64 rows (windows) and 6 columns (features), one matrix data is based on the approximations and the other is given by the details coefficients. Each window is represented by six features of the wavelets coefficients, namely Energy, Mean, Variance, Skewness, Kurtosis, and Coefficient of Variation (CV).

Next, the principal component analysis is conducted on the two data matrices, and the number of principles components are selected based on the explained variances. We selected a number of principal components that explain at least 90% and at most 95% of the data. Then, the Hotelling $T^2_{a_j}$ and $T^2_{d_j}$ are calculated into the reduced space, see Sect. 8.2.2, for the approximation and details coefficients, respectively. Finally, $T^2 = \sqrt{T^2_{a_j} + T^2_{d_j}}$ statistic is then calculated to represent the original MS spectrum, see Fig. 8.6. Each MS sample will be given by T^2 statistic, and the classification model will be built on T^2 statistics for each patient.

8.3.1 Results and Performance

The classification model is built in two phases, a training phase, and a test phase. 80% of the dataset is used as the training dataset (78 Cancer, 46 Normal) and 20%

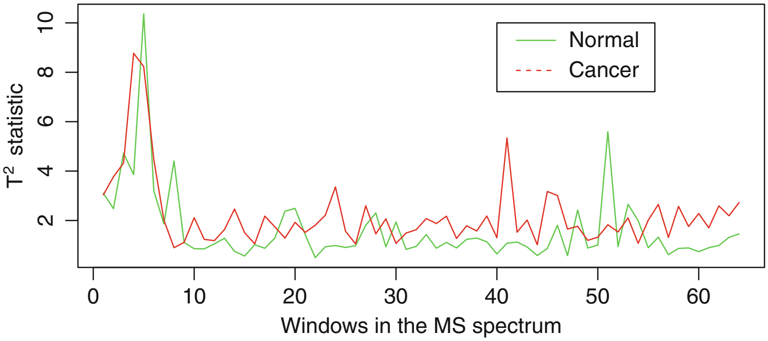

Fig. 8.6 T^2 statistic for normal and cancer MS spectrum

as the testing data set. The MS data are then classified as normal or cancer. The classification results will be given in terms of accuracy, sensitivity, and specificity, as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{8.16}$$

$$Sensitivity = \frac{TP}{TP + FN}, \tag{8.17}$$

$$Specificity = \frac{TN}{TN + FP}, \tag{8.18}$$

where TP is True Positive, TN True Negative, FP False Positive, and FN False Negative. The proposed framework achieves a reasonable classification performance. The combination of wavelets coefficients and higher order statistics provide useful discriminatory information about MS data.

The classification model is developed using the training set and then tested using the testing samples. The predictive procedure optimizes the model parameters to build a model that fits the training data as well as possible, which then may lead to an overfitting. Cross-validation can be used as a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set, therefore, this will help overcome the overfitting problem.

We performed a k-fold cross-validation procedure in order to avoid overfitting and evaluate the generalization capabilities of our model. The parameter k varies into {5, 10, 20, 50}. A Monte-Carlo simulation repeating the whole process including the selection of the training and testing sets is also performed. 100 simulation runs were conducted. Consequently, the results reported are the averages and standard errors of the performance measures given in Eqs. (8.16)–(8.18).

Table 8.1 shows a summary of the performance results of our proposed method. The accuracy classification is 100% on average with 0 standard error. The average sensitivity and specificity are equal to 100% for the training set and the testing set.

Table 8.1 The average (standard error) of the performance results using 100 replications of k-Fold cross validation, and the hyperparameters C = 2 and γ = 0.0009765 obtained from the grid search optimization

k ↓	Accuracy(SE) %	Sensitivity		Specificity	
		Training set	Testing set	Training set	Testing set
5	99.91 (0.12)	1 (0)	1 (0)	1 (0)	1 (0)
10	100 (0)	1 (0)	1 (0)	1 (0)	1 (0)
20	100 (0)	1 (0)	1 (0)	1 (0)	1 (0)
50	100 (0)	1 (0)	1 (0)	1 (0)	1 (0)

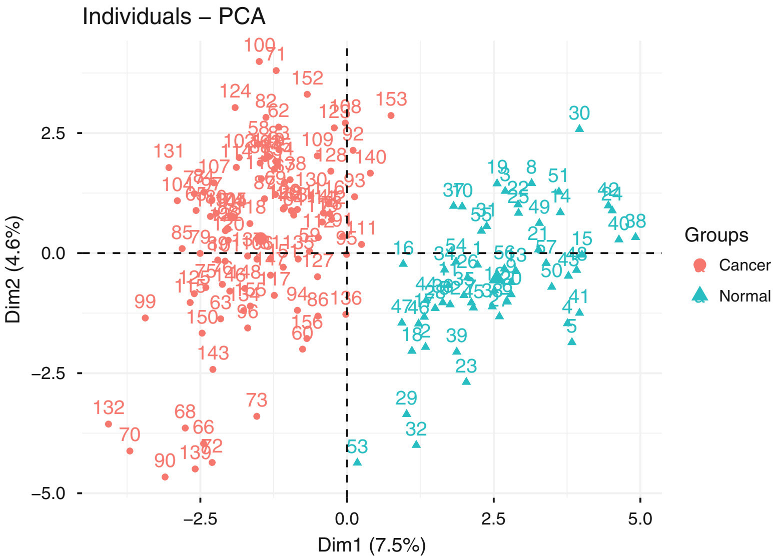

Fig. 8.7 The first two principal components on T^2 statistic for 99 cancer samples and 57 normal samples

In order to investigate the excellent performance of the proposed method, we conducted a principal component analysis (PCA) on the 156 samples of MS data (57 normal and 99 cancer). The PCA is applied to a data matrix of 156 rows (patients) and 64 columns (T^2 statistics). The results given in Fig. 8.7 show clearly that the proposed method ingeniously separates the cancer MS samples from the Normal MS samples.

This result has a valuable scientific impact on public health, especially on the early cancer detection. In fact, automatic classification of these mass spectrometry patterns will definitely help physicians in the diagnosis of diseases such as cancer. In addition, the higher classification performance we have, the more confident in the diagnosis we become.

8.4 Conclusion

The MS data are important for clinical diagnosis and health advances. Preprocessing methods and transformation such as wavelets analysis and principal component analysis can help face high dimensionality and reduce noise.The accuracy of the proposed model is 100% on average with 0 standard error. The average sensitivity and specificity are equal to 100% for the training set and the testing set. This paper contributes to the development of accurate models for MS classification. We aim at applying the proposed method to other MS data of different types of cancer.

References

Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature, 422*(6928), 198–207.

Awedat, K., Abdel-Qader, I., & Springstead, J. R. (2016). Mass spectrometry sensing data for robust cancer classification. In *Electro Information Technology (EIT), 2016 IEEE International Conference on* (pp. 0258–0262). Piscataway: IEEE.

Cohen, A., Daubechies, I., & Feauveau, J.-C. (1992). Biorthogonal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics, 45*(5), 485–560.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning, 20*(3), 273–297.

Cravatt, B. F., Simon, G. M., & Yates Iii, J. R. (2007). The biological impact of mass-spectrometry-based proteomics. *Nature, 450*(7172), 991.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. In *ICML* (Vol. 1, pp. 74–81).

Daubechies, I. (1992). *Ten lectures on wavelets*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

de Noo, M. E., Mertens, B. J., Özalp, A., Bladergroen, M. R., van der Werff, M. P., van de Velde, C. J., et al. (2006). Detection of colorectal cancer using maldi-tof serum protein profiling. *European Journal of Cancer, 42*(8), 1068–1076.

Diamandis, E. P. (2004). Mass spectrometry as a diagnostic and a cancer biomarker discovery tool opportunities and potential limitations. *Molecular & Cellular Proteomics, 3*(4), 367–378.

Donoho, D. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory, 41*(3), 613–627.

Donoho, D. L., & Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika, 81*(3), 425–455.

Donoho, D. L., & Johnstone, J. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Jouranl of the American Statistical Association, 90*, 1200–1224.

Du, J., Wu, X.-M., Wang, B., Su, H.-J., Ma, K., & Zhang, H.-Q. (2009). Wavelet transform and bagging predictor approaches to cancer identification from mass spectrometry-based proteomic data. In *Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009. 3rd International Conference on* (pp. 1–4). Piscataway: IEEE.

Dubitzky, W., Granzow, M., & Berrar, D. P. (2007). *Fundamentals of data mining in genomics and proteomics*. Berlin: Springer Science and Business Media.

Gao, H.-Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *Journal of Computational and Graphical Statistics, 7*(4), 469–488.

Gao, H.-Y., & Bruce, A. G. (1997). Waveshrink with firm shrinkage. *Statistica Sinica, 7*(4), 855–874.

Ge, G., & Wong, G. W. (2008). Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles. *BMC Bioinformatics, 9*(1), 275.

Gromski, P. S., Xu, Y., Correa, E., Ellis, D. I., Turner, M. L., & Goodacre, R. (2014). A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. *Analytica Chimica Acta, 829*, 1–8.

Jolliffe, I. T. (1986). Principal component analysis and factor analysis. In *Principal component analysis* (pp. 115–128). Berlin: Springer.

Lancashire, L. J., Lemetre, C., & Ball, G. R. (2009). An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies. *Briefings in Bioinformatics, 10*, 315–329. https://doi.org/10.1093/bib/bbp012.

Li, Y., & Zeng, X. (2016). Serum seldi-tof ms analysis model applied to benign and malignant ovarian tumor identification. *Analytical Methods, 8*(1), 183–188.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*(7), 674–693.

Misiti, M., Misiti, Y., Oppenheim, G., & Poggi, J. (1996). *Wavelet toolbox*. Natick, MA: The MathWorks Inc.

Morris, J. S., Brown, P. J., Baggerly, K. A., & Coombes, K. R. (2006). Analysis of mass spectrometry data using bayesian wavelet-based functional mixed models. In *Bayesian inference for gene expression and proteomics* (pp. 269–288). Cambridge: Cambridge University Press.

Nguyen, T., Nahavandi, S., Creighton, D., & Khosravi, A. (2015). Mass spectrometry cancer data classification using wavelets and genetic algorithm. *FEBS Letters, 589*(24), 3879–3886.

Ohn, S.-Y., Chi, S.-D., & Heo, C. (2016). Identification of breast cancer by classification of proteome patterns. *International Journal of Modeling, Simulation, and Scientific Computing, 7*(04), 1643004.

P. Datasets for Breast Cancer (2004). http://bioinformatics.mdanderson.org/pubdata.html.

Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., et al. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet, 359*(9306), 572–577.

Schleif, F.-M., Lindemann, M., Diaz, M., Maaß, P., Decker, J., Elssner, T., et al. (2009). Support vector classification of proteomic profile spectra based on feature extraction with the bi-orthogonal discrete wavelet transform. *Computing and Visualization in Science, 12*(4), 189–199.

Sharma, A., & Singh, S. (2016). Neural network for diagnosis of ovarian cancer based on proteomic patterns in serum. *Journal of Scientific and Technical Advancements, 2*(2), 25–27.

Shawe-Taylor, J. & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press.

Tang, K.-L., Li, T.-H., Xiong, W.-W., & Chen, K. (2010). Ovarian cancer classification based on dimensionality reduction for seldi-tof data. *BMC Bioinformatics, 11*(1), 109.

Vapnik, V. (2013). *The nature of statistical learning theory*. Berlin: Springer Science and Business Media.

Ward, D. G., Nyangoma, S., Joy, H., Hamilton, E., Wei, W., Tselepis, C., et al. (2008). Proteomic profiling of urine for the detection of colon cancer. *Proteome Science, 6*(1), 19.

Wu, J., Ji, Y., Zhao, L., Ji, M., Ye, Z., & Li, S. (2016). A mass spectrometric analysis method based on ppca and svm for early detection of ovarian cancer. *Computational and Mathematical Methods in Medicine, 2016*, 6169249.

Yildiz, P. B., Shyr, Y., Rahman, J. S., Wardwell, N. R., Zimmerman,L. J., Shakhtour, B., et al. (2007). Diagnostic accuracy of maldi mass spectrometric analysis of unfractionated serum in lung cancer. *Journal of Thoracic Oncology, 2*(10), 893–901.

Yu, J., Ongarello, S., Fiedler, R., Chen, X., Toffolo, G., Cobelli, C., et al. (2005). Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics, 21*(10), 2200–2209.