Journal of Computer Science 9 (8): 1061-1068, 2013
ISSN: 1549-3636

doi:10.3844/jcssp.2013.1061.1068 Published Online 9 (8) 2013 (http://www.thescipub.com/jcs.toc)

SPECTRAL DOMAIN FEATURES FOR OVARIAN CANCER DATA ANALYSIS

[1]Ahmed Farag Seddik, [1]Riham Amin Hassan and [2]Mahmoud A. Fakhreldein

[1]Department of Biomedical, Faculty of Engineering, Helwan University, Cairo, Egypt
[2]Department of Computers and Systems, Electronics Research Institute, Cairo, Egypt

Received 2013-04-30, Revised 2013-05-23; Accepted 2013-07-09

ABSTRACT

The early detection of cancer is crucial for successful treatment. Medical researchers have investigated a number of early-diagnosis techniques. Recently, they have discovered that some cancers affect the concentration of certain molecules in the blood, which allows early diagnosis by analyzing the blood mass spectrum. Researchers have developed several techniques for the analysis of the mass-spectrum curve analysis and used them for the detection of prostate, ovarian, breast, bladder, pancreatic, kidney, liver and colon cancers. In this study we propose a new technique that uses the spectral domain features such as wavelet transform and Fourier transform for the analysis of the ovarian cancer data to differentiate between normal and patients with malignant cancer. We used two different classifiers for the original data, the first one is a feed forward artificial neural network classifier which gave a sensitivity of 96%, specificity of 88% and accuracy of 94%. The second used classifier is the linear discriminant analysis classifier which separated the cancer from healthy samples with sensitivity of 79%, specificity of 75% and accuracy of about 81%. After transforming the data to the spectral domain using the Fourier transform the performance was degraded. The experimental results showed that the performance of the wavelet transform based system was superior to other techniques as it gave a sensitivity of 98%, specificity of 96% and accuracy of 95%.

Keywords: Spectral Domain Features, Cancer Data, Surface-Enhanced Laser Desorption and Ionization

1. INTRODUCTION

Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary. Profile patterns are generated using Surface-Enhanced Laser Desorption and Ionization (SELDI) protein mass spectrometry (**Fig. 1**). This technology has the potential to improve clinical diagnostics tests for cancer pathologies. The goal is to select a reduced set of measurements or “features” that can be used to distinguish between cancer and control patients. These features will be ion intensity levels at specific mass/charge values.

The blood mass spectrum is a curve (**Fig. 2**), where the x-axis shows the ratio of the weight of a specific molecule to its electric charge and the y-axis is the signal intensity for the same molecule. The mass-spectrum analysis is a fast inexpensive procedure based on a sample of a patient’s blood and it may potentially allow cancer screening with little discomfort to a patient (Bakhtiar and Nelson, 2001; Bakhtiar and Tse, 2000; Yates, 2000).

In this study we propose a new technique that uses the spectral domain features such as wavelet transform and Fourier transform for the analysis of the ovarian cancer data to differentiate between normal and patients with malignant cancer. When feed forward artificial neural network classifier is compared with LDA classifier, it gives more efficiency than LDA but it gives less efficiency when wrapping curve is performed.

Corresponding Author: Ahmed Farag Seddik, Department of Biomedical, Faculty of Engineering, Helwan University, Cairo, Egypt

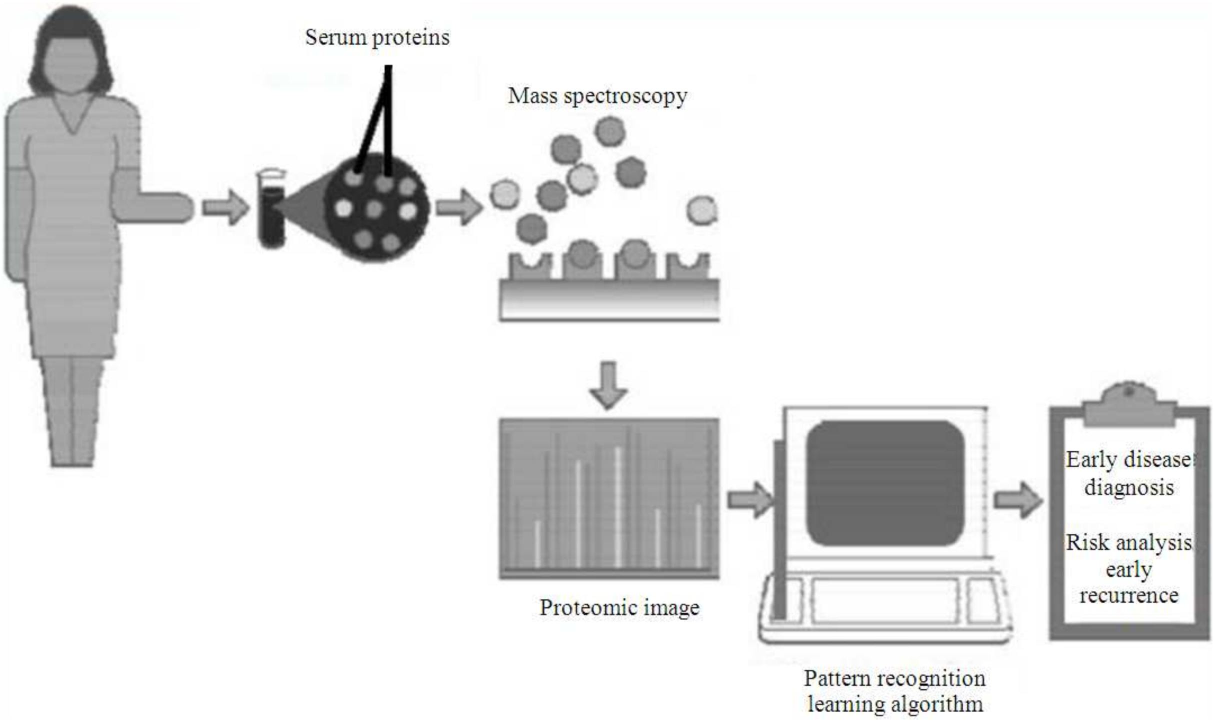

Fig. 1. Surface-enhanced laser desorption and ionization (SELDI) protein mass spectrometry

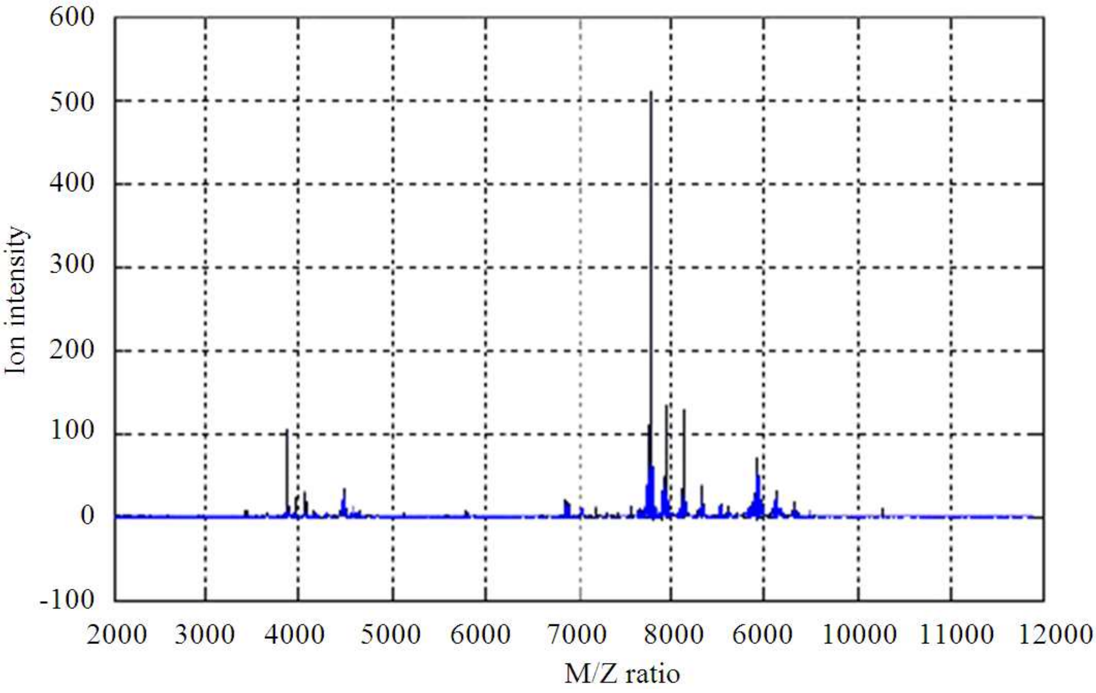

Fig. 2. The blood mass spectrum

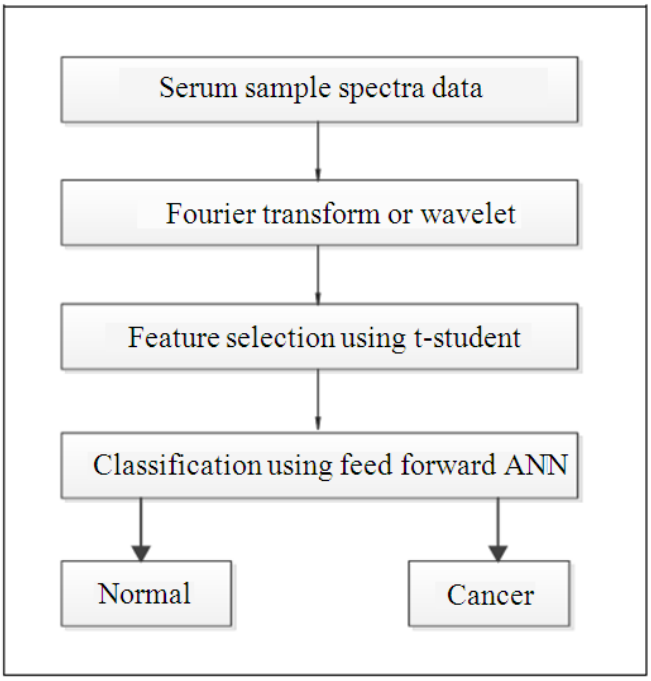
Serum sample spectra data
Fourier transform or wavelet
Feature selection using t-student
Classification using feed forward ANN
Normal
Cancer

Fig. 3. Simple block diagram for the first technique

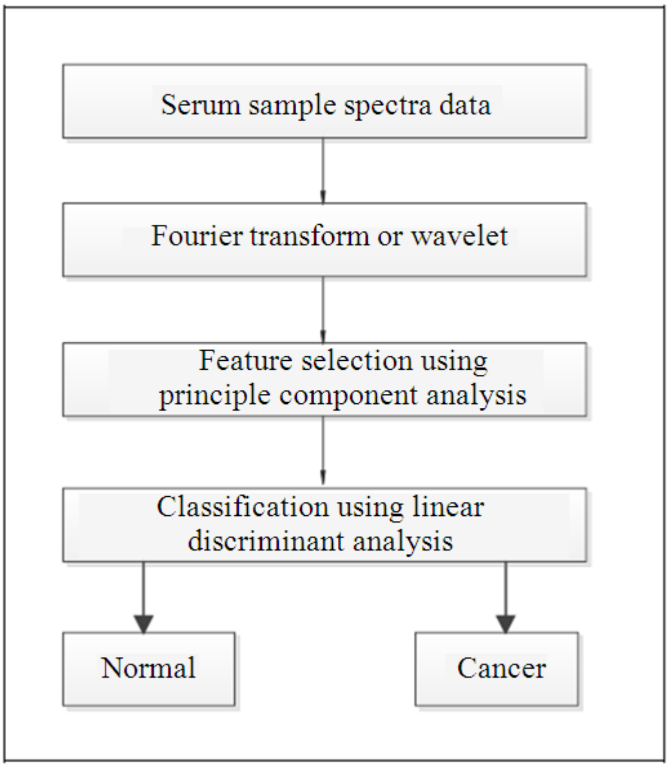
Serum sample spectra data
Fourier transform or wavelet
Feature selection using principle component analysis
Classification using linear discriminant analysis
Normal
Cancer

Fig. 4. Simple block diagram for the second technique

1.1. Literature Review

Medical researchers have developed techniques for the detection of early cancer based on protein markers, which are certain molecules in body tissues and fluids (Poon and Johnson, 2001), but these techniques are often inaccurate. For example, the specificity of an antigen method for the prostate-cancer detection is only 25-30%, although its sensitivity is high (Adam *et al*., 2001); as another example, the sensitivity of a similar method for breast cancer is 23% and its specificity is 69% (Li *et al*., 2002). Recently, researchers have developed a new cancer-detection method based on the application of data mining to the mass spectra of patients' tissue cells, blood, serum and other body fluids (Petricoin and Liotta, 2002; Petricoin *et al*., 2002c; Wulfkuhle *et al*., 2003).

In previous reports, researchers have compared results obtained with several well-known classification methods to distinguish ovarian cancer patients from normal individuals based on MS data obtained on serum samples. Overall, they have found that the Random Forest (RF) (Wu *et al*., 2003) approach both leads to an overall lower misclassification rate as well as to a more stable assessment of classification errors. Therefore, their preliminary analyses suggest that RF and methods similar in nature to RF may be more useful than other methods to classify samples based on MS data. Compared to Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) methods (Huang *et al*., 2012), RF has the advantage of not requiring the number of variables used to be less than the number of subjects in the study, which is a clear advantage for the analysis of MS data as the number of m/z versus intensity data points is very large. In addition, RF is able to handle interactions among variables.

In this study we proposed the use of feed forward artificial neural network as a classifier and compared it with the LDA classifier. The proposed technique gives higher classification performance than LDA.

1.2. The Proposed Technique

In this study, we used two techniques to reduce and classify the extracted serum sample spectrum data. In the first technique, we applied Fourier transform or wavelet transform (**Fig. 3**) on the data. Then we used the student t-test to select the features and perform classification using feed forward neural network. But in the second technique, we used the principle component analysis to select the features and perform classification using linear discriminant analysis (**Fig. 4**).

2. MATERIALS AND METHODS

The ovarian cancer dataset in this study comes from the FDA-NCI Clinical Proteomics Program Databank. This study uses the high-resolution ovarian cancer data set that was generated using the WCX2 protein array. The sample set includes 95 controls and 121 ovarian cancers. An extensive description of this data set and excellent introduction to this promising technology can be found in (Conrads *et al*., 2004a; Petricoin *et al*., 2002a). The dataset includes three matrices as shown in **Table 1**.

Each column in Y represents measurements taken from a patient. There are 216 columns in Y representing 216 patients, out of which 121 are ovarian cancer patients and 95 are normal patients. Each row in Y represents the ion intensity level at a specific mass-charge value indicated in MZ. There are 15000 mass-charge values in MZ and each row in Y represents the ion-intensity levels of the patients at that particular mass-charge value. The variable grp holds the index information as to which of these samples represent cancer patients and which ones represent normal patients. An extensive description of this data set and excellent introduction to this promising technology can be found in (Conrads *et al*., 2004b; Petricoin *et al*., 2002b).

2.1. Feature selection and Ranking

This is a typical classification problem in which the number of features is much larger than the number of observations, but in which no single feature achieves a correct classification, therefore we need to find a classifier which appropriately learns how to weight multiple features and at the same time produce a generalized mapping which is not over-fitted.

2.2. Student T-Test Method

A simple approach for finding significant features is to assume that each M/Z value is independent and compute a two-way t-test (Sawilowsky, 2005). We used the student t-test method to rank the features and we got an index to the most significant M/Z values ranked by the absolute value of the t-test statistic value. This feature selection method is also known as a filtering method, where the learning algorithm is not involved on how the features are selected. In this study we selected the top 200 features based on the t-test value.

2.3. Principal Component Analysis (PCA)

It is a way of identifying patterns in data and expressing the data in such a way as to highlight their similarities and differences (Abdi and Williams, 2010). Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data. In this study we used PCA as a feature reduction technique to reduce the high dimensionality of feature space to only 200 features.

2.4. Classification

After selection of the most 200 significant features using one of the feature selection techniques described above, we used this information to classify the cancer and normal samples.

2.5. Using a Feed Forward Neural Network

A neural network is a massively parallel distributed processor made up of simple processing units that have a natural tendency for storing experiential knowledge and making it available for us. Artificial Neural Network (ANN) is a type of artificial intelligence technique that mimics the behavior of the human brain (Hu and Hwang, 2010).

First, the data is separated into inputs and targets. The significant features identified will act as the inputs to the neural network. The targets for the neural network will be the logical indices of cancer samples. Cancer samples will hence be identified with 1’s and normal samples will be identified with 0’s.

A 1-hidden layer feed forward neural network with 5 hidden layer neurons is created and trained. The 216 input and target patterns are divided into 60% training, 20% validation and 20% for testing. The training set is used to teach the network. Training continues as long as the network continues improving on the validation set. The test set provides a completely independent measure of network accuracy.

The trained neural network were tested with the testing samples we partitioned from the main dataset. The testing data was not used in training in any way and hence provides an “out-of-sample” dataset to test the network. This gave us a sense of how well the network will do when tested with data from the real world.

Table 1. Dataset

Data matrix name	Size
MZ	15000×1
Y	15000×216
grp	216×1

2.6. Using Linear Discriminant Analysis (LDA)

Linear discriminant analysis (Hastie *et al*., 2005) has been explored for the probabilistic classification of healthy versus ovarian cancer serum samples using proteomics data from Mass Spectrometry (MS).The linear discriminant analysis method consists of searching, some linear combinations of selected variables, which provide the best separation between the considered classes.

3. RESULTS AND DISCUSSION

Clinicians use three standard measures of the effectiveness of diagnosis techniques: sensitivity, specificity and accuracy. The sensitivity is the probability of the correct diagnosis for a patient with cancer, the specificity is the chances of the correct diagnosis for a healthy person and the accuracy is the chances of the correct diagnosis for the overall population of healthy and sick people.

Table 2 depicts that after transforming the data to the spectral domain using the Fourier transform the performance was degraded. From **Table 2**, it is clearly noticed that the performance of the wavelet transform based system was superior to other techniques as it gave a sensitivity of 98%, specificity of 96% and accuracy of 95%.

Also, we can notice that feed forward neural network classifier gives better specificity, sensitivity and accuracy than LDA.

The spectral transformation for the data using Fourier transform degraded the accuracy and sensitivity while maintaining the specificity. On the other hand, the wavelet transform increased the performance of the feed forward artificial neural network classifier. Although the performance of the feed forward neural network is better than LDA the wrapping curve of LDA (**Fig. 8**) is more accurate than the feed forward neural network that is because training multiple times will generate different results due to different initial conditions and sampling (**Fig. 5-7**).

Table 2. Comparison between the different methods

Technique	Specificity (%)	Sensitivity (%)	Accuracy (%)
Neural network	88	96	94.0
Neural network with fourier transform	88	89	85.0
Neural network with Wavelet transform	98	96	95 .0
LDA	75	79	80.7

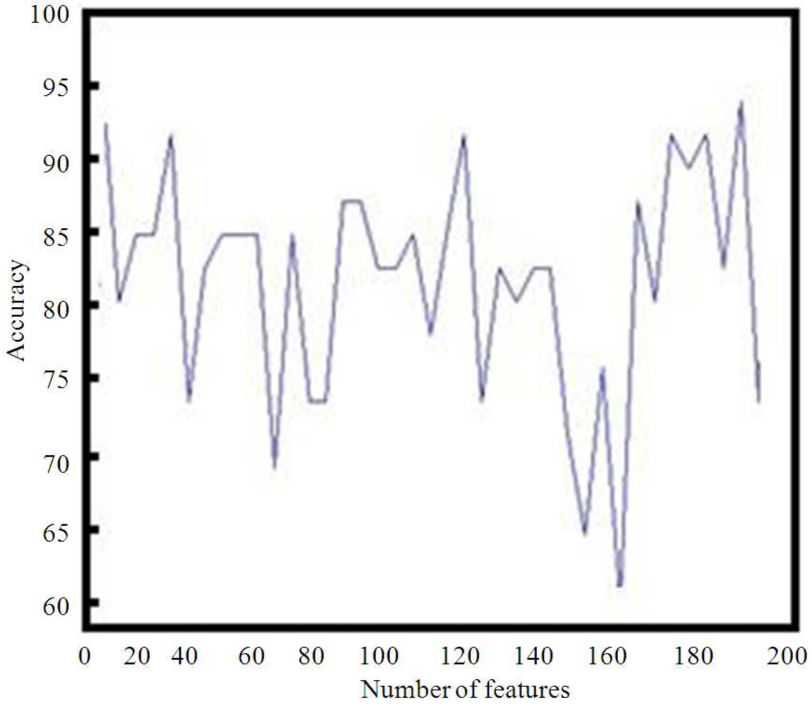

Fig. 5. Wrapping curve in feed forward neural network classifier

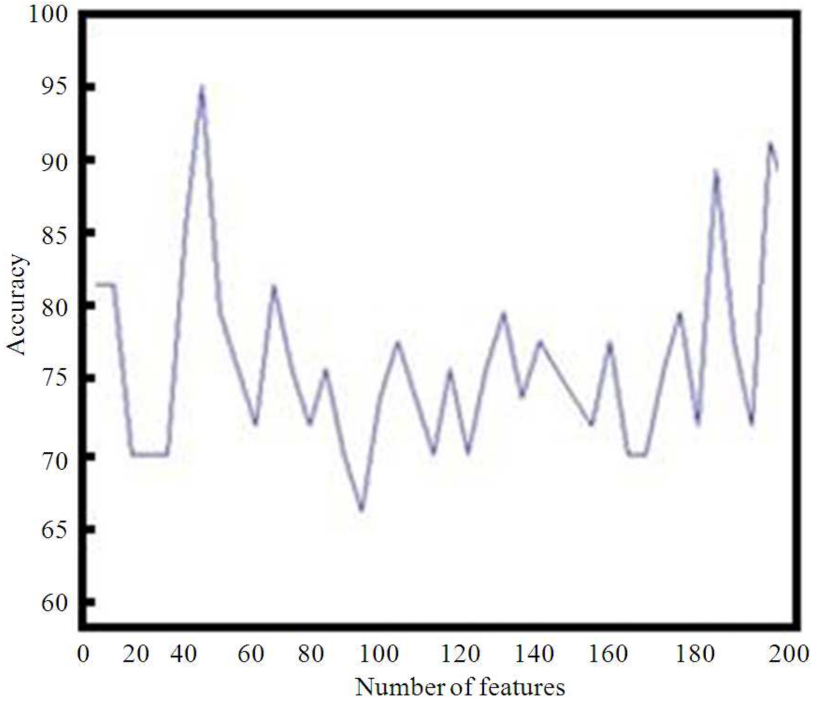

Fig. 6. Wrapping curve in feed forward neural network classifier with wavelet transform of data

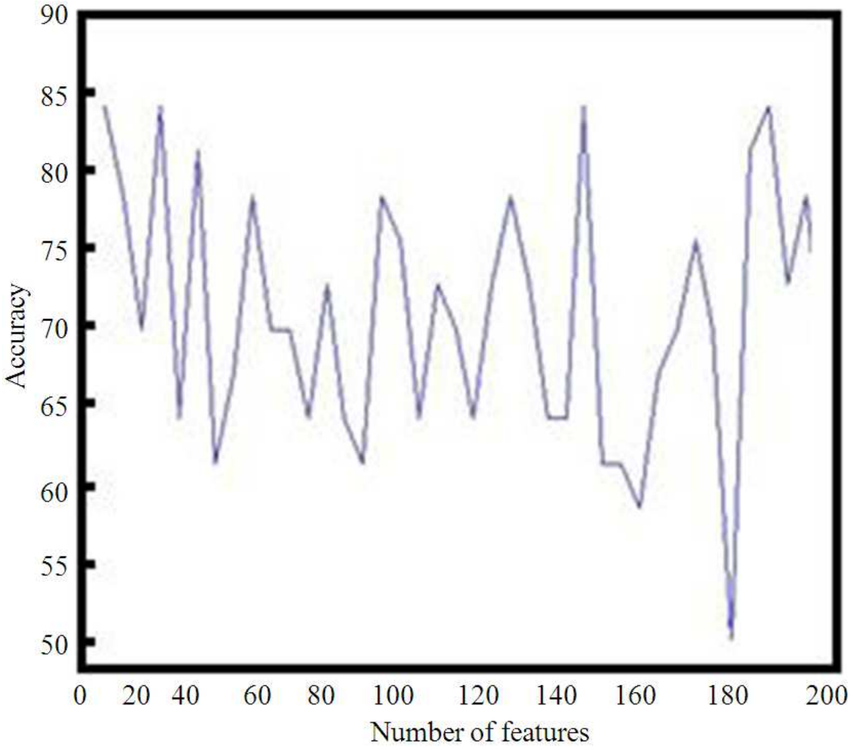

Fig. 7. Wrapping curve in feed forward neural network classifier with fourier transform of data

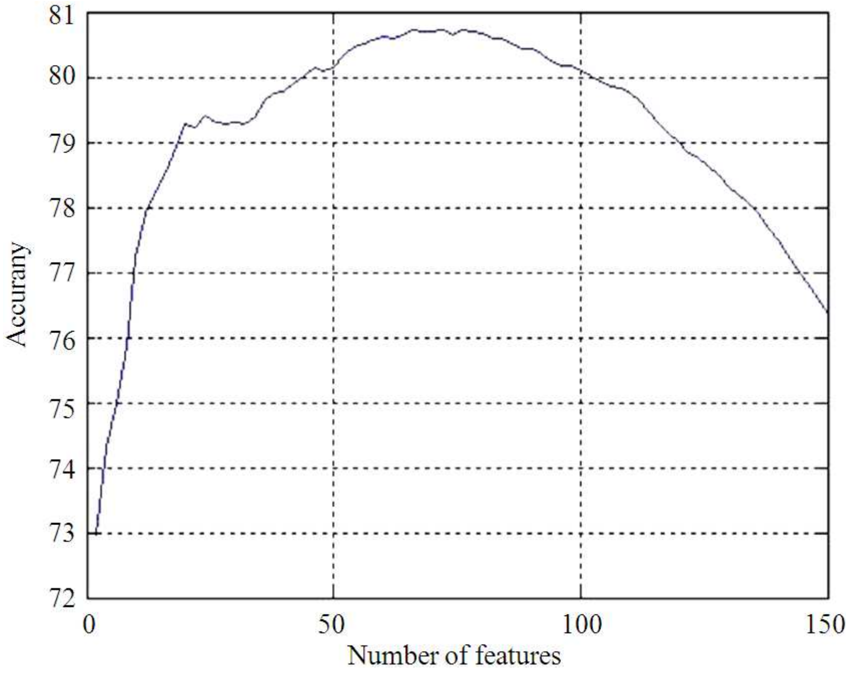

Fig. 8. Wrapping curve in linear discriminant analysis classifier

4. CONCLUSION

In this study we proposed a new methodology to distinguish between the normal patient and cancer based on the spectral domain features of the data. Experimental results have demonstrated that the feed forward neural network classifier gives better specificity, sensitivity and accuracy than LDA and after transforming the data to the spectral domain using the Fourier transform the accuracy and sensitivity were degraded while maintaining the specificity. On the other hand, the wavelet transform increased the performance of the feed forward artificial neural network classifier.

The amount and quality of the data are key components of the diagnostic accuracy. The measuring process may contain many features that create problems for the data mining techniques. The datasets could be consisted of a large volume of heterogeneous data fields which usually complicates the use of data mining techniques.

5. REFERENCES

Abdi, H. and L.J. Williams, 2010. Principal component analysis. WIREs Comp. Stat., 2: 433-459. DOI: 10.1002/wics.101

Adam, B.L., A. Vlahou, O.J. Semmes and G.L. Wright, 2001. Proteomic approaches to biomarker discovery in prostate and bladder cancers. Proteomics, 1: 1264-1270. PMID: 11721637

Bakhtiar, R. and F.L.S. Tse, 2000. Biological mass spectrometry: A primer. Mutagenesis, 15: 415-430. PMID: 10970448

Bakhtiar, R. and R.W. Nelson, 2001. Mass spectrometry of the proteome. Mol. Pharmacol., 60: 405-415.

Conrads, T.P., V.A. Fusaro, S. Ross, D. Johann and V. Rajapakse *et al*., 2004a. High-resolution serum proteomic features for ovarian cancer detection. Endoc. Relat. Cancer, 11: 163-178. PMID: 15163296

Conrads, T.P., V.A. Fusaro, S. Ross, D. Johann and V. Rajapakse, *et al*., 2004b. High-resolution serum proteomic features for ovarian cancer detection. Endocr. Relat. Cancer, 11: 163-178. PMID: 15163296

Hastie, T., R. Tibshirani and J. Friedman, 2005. The Elements of Statistical Learning: Data Mining, Inference and Prediction. 2nd Edn., Springer, New York, ISBN-10: 0387848584, pp: 745.

Hu, Y.H. and J.N. Hwang, 2010. Handbook of Neural Network Signal Processin. 1st Edn., Taylor and Francis, Boca Raton, ISBN-10: 0849323592, pp: 408.

Huang, H., Y. Liu, L. Bosch, B. Cranen and L. Boves, 2012. Knowledge-based quadratic discriminant analysis. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 25-30, IEEE Xplore Press, Kyoto, pp: 4145-4148. DOI: 10.1109/ICASSP.2012.6288831

Li, J., Z. Zhang, J. Rosenzweig, Y.Y. Wang and D.W. Chan, 2002. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. Clin. Chem., 48:1296-1304. PMID: 12142387

Petricoin, E.F. and L.A. Liotta, 2002. Proteomic analysis at the bedside: Early detection of cancer. Trends Biotechnol., 20: S30-S34. DOI: 10.1016/S1471-1931(02)00204-5

Petricoin, E.F., A.M. Ardekani, B.A. Hitt, P.J. Levine and V.A. Fusaro *et al*., 2002a. Use of proteomic patterns in serum to identify ovarian cancer. Lancet, 359: 572-577. PMID: 11867112

Petricoin, E.F., A.M. Ardekani, B.A. Hitt, P.J. Levine and V.A. Fusaro *et al*., 2002b. Use of proteomic patterns in serum to identify ovarian cancer. Lancet, 359: 572-577. PMID: 11867112

Petricoin, E.F., K.C. Zoon, E.C. Kohn, J.C. Barrett and L.A. Liotta, 2002c. Clinical proteomics: Translating benchside promise into bedside reality. Nature Rev. Drug Disco., 1: 683-695. PMID: 12209149

Poon, T.C.W. and P.J. Johnson, 2001. Proteome analysis and its impact on the discovery of serological tumor markers. Clin. Chim. Acta, 313: 231-239. DOI: 10.1016/S0009-8981(01)00677-5

Sawilowsky, S.S., 2005. Misconceptions leading to choosing the t-test over the Wilcoxon Mann-Whitney U test for shift in location parameter. J. Mod. Applied Stat. Meth., 4: 598-600.

Wu, B., T. Abbott, D. Fishman, W. McMurray and G. Mor, 2003. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics, 19: 1636-1643. DOI: 10.1093/bioinformatics/btg210

Wulfkuhle, J.D., L.A. Liotta and E.F. Petricoin, 2003. Proteomic applications for the early detection of cancer. Nat. Rev. Cancer, 3: 267-275. PMID: 12671665

Yates, J.R., 2000. Mass spectrometry. From genomics to proteomics. Trends Genet., 16: 5-8. PMID: 10637622