

# **Proteomic MS-spectra decomposition into intensity-regions for identifying potential biomarkers and improving discrimination accuracy between normal and cancerous spectra**

**Panagiotis Bougioukos<sup>1</sup>, Dionisis Cavouras<sup>2</sup>, Dimitris Glotsos<sup>2</sup>,  
Ioannis Kalatzis<sup>2</sup>, George Nikiforidis<sup>1</sup>, and Anastasios Bezerianos<sup>1</sup>**

cavouras@teiath.gr

**Department of Medical Physics, School of Medicine, University of Patras, Greece  
Medical Signal and Image Processing Lab, Department of Medical Instruments Technology,  
Technological Educational Institute of Athens, Greece**

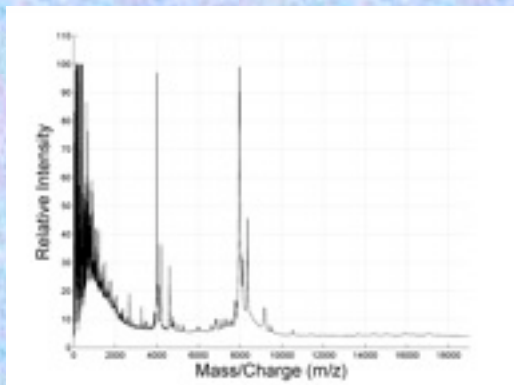
# MOTIVATION

- Biologists' interest [*Morris S. J., Bioinformatics, 21, 2005*] in investigating the usefulness of the low intensity peaks. Thus,
- Peaks at various intensity levels (Peak Intensity Zones) were examined for their capability in discriminating healthy from cancer cases

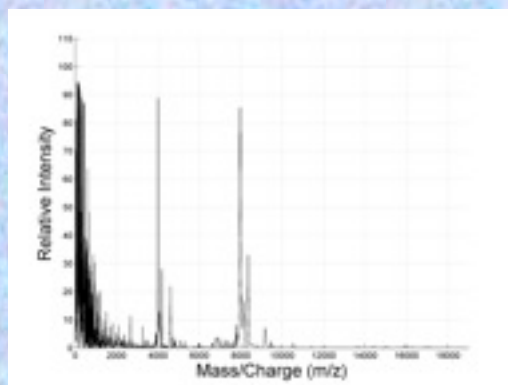
# AIM

- **robust extraction of biological meaningful mass spectral peaks for**
  - a/ determination of potential meaningful cancer biomarkers ( $m/z$  values), extracted from different MS-spectra intensity regions
  - b/ reliable and effective separation of normal from cancerous MS-spectra

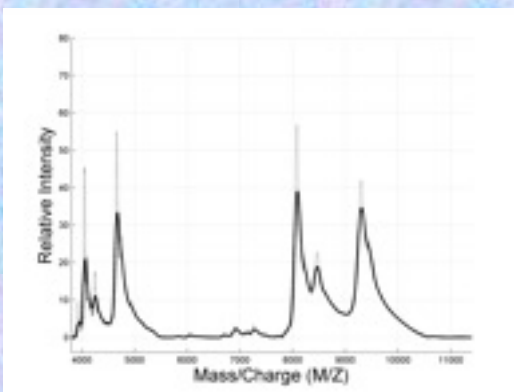
# METHODS (I): Baseline Subtraction-Normalization-Smoothing



**Fig. a:** Original MS-spectrum



**Fig. b:** MS-spectrum with baseline correction



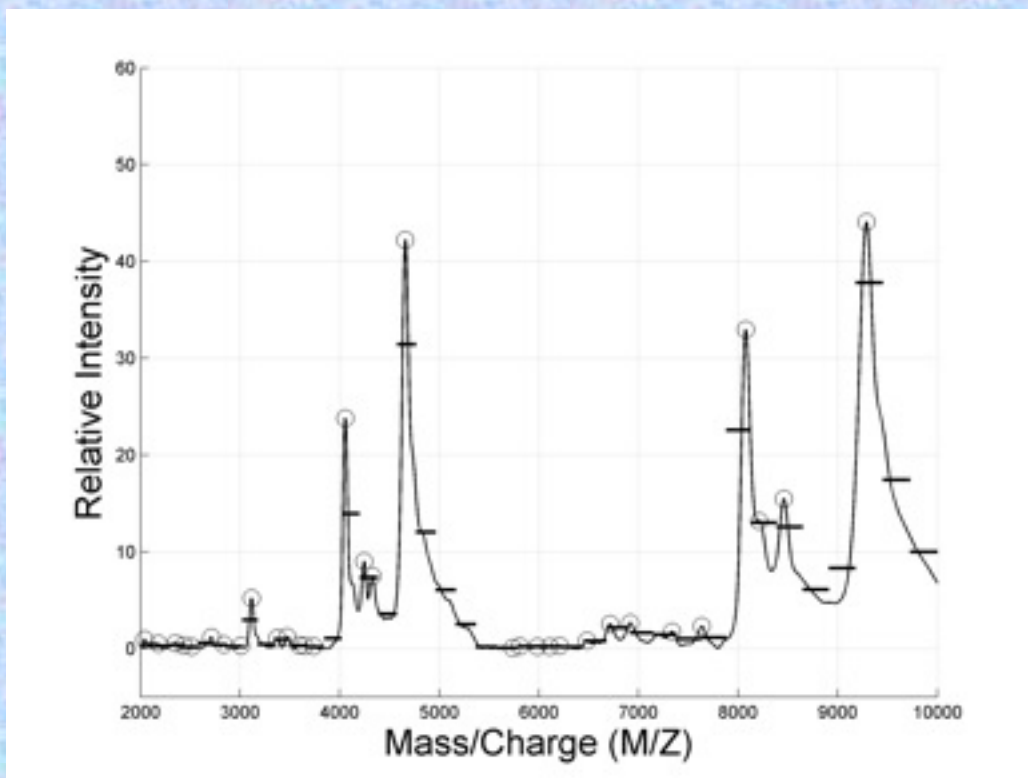
**Fig. c:** A region of the original (gray line) and the smoothed (bold line) MS-spectra.

**Baseline Subtraction:** MS-spectra baseline drift (due to chemical and electronic noise) was estimated [L. Andrade, et al., VLSI, 35, 2003], by using multiple shifted windows of 200 bins ( $m/z$  values) size, and spline approximation was used to regress the varying baseline, which was subtracted from the spectrum (see Fig (a) & (b) ).

**MS-spectra Normalization:** to reduce variation in signal intensity between MS-spectra

**MS-spectra Smoothing:** for reducing the spectrum's spikes that appear to constitute peaks which are not replicates at all spectra using the Lowess [Cleveland, W. S., J Amer Statist Assoc. 1979] smoothing technique (Fig. (c) ).

## METHODS (II): Noise estimation

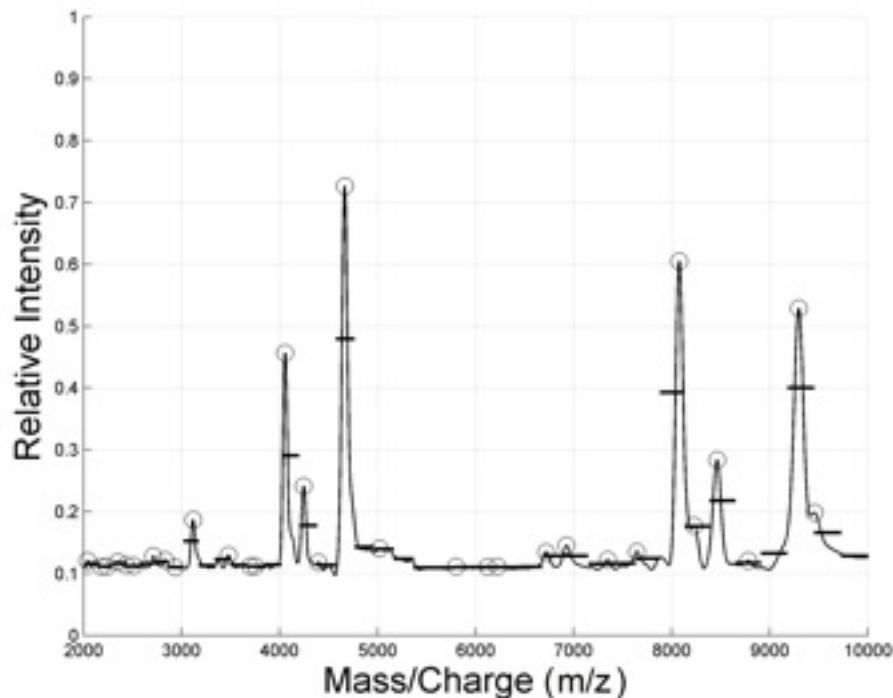


A region of the smoothed spectrum, showing the estimated noise levels (horizontal lines) and peaks (circle symbol)

**Noise estimation:** Determines the intensity value above which the spectrum's data points are considered as informative [G. Bhanot Proteomics 6, 2006]

Local noise was computed according to  $\text{Noise} = \text{mean} + \text{std}$  at each position of a sliding window, considering only intensity values below the the 90th percentile of the local histogram [X. Wang, Proteomics 6, 2006]. MS-Spectral points below the local\_noise level were considered as noise and were omitted from further analysis. The width of the sliding window was fixed, comprising 1% of all spectral points, whereas the size of the sliding window was variable, since the distance between adjacent spectral points is not equal.

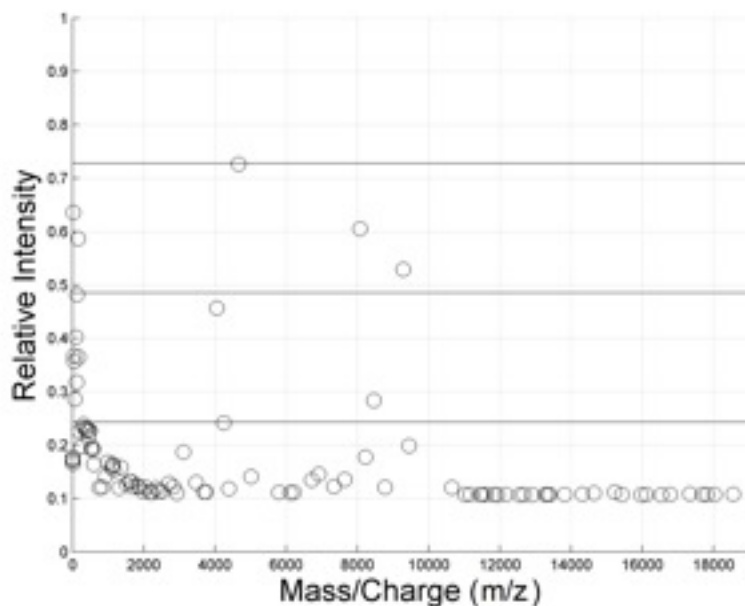
## METHODS (III): Peak detection



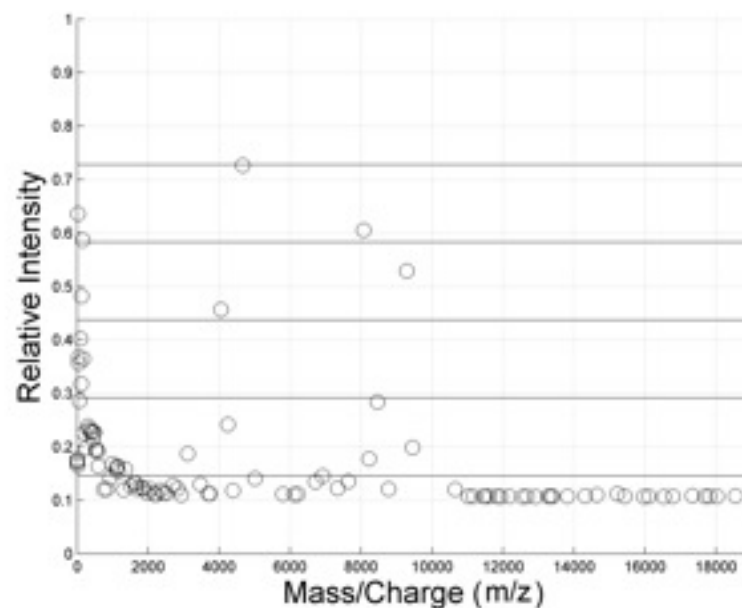
Peaks were detected using a differentiation method [K.R. Coombes, *Proteomics* 5, 2005]. Selected peaks were considered as encoding information concerning potential biomarkers and were used for further analysis [J. S. Morris *Bioinformatics*, 21, 2005].

## METHODS (IV): Peak intensity group generation

- The low intensity peaks are frequently the peaks in which biologists are focused in order to extract potentially biological conclusions [*Morris S. J., Bioinformatics, 21, 2005*]
- Proposed method: Each MS-spectrum's peaks were assigned into 3 or 5 peak

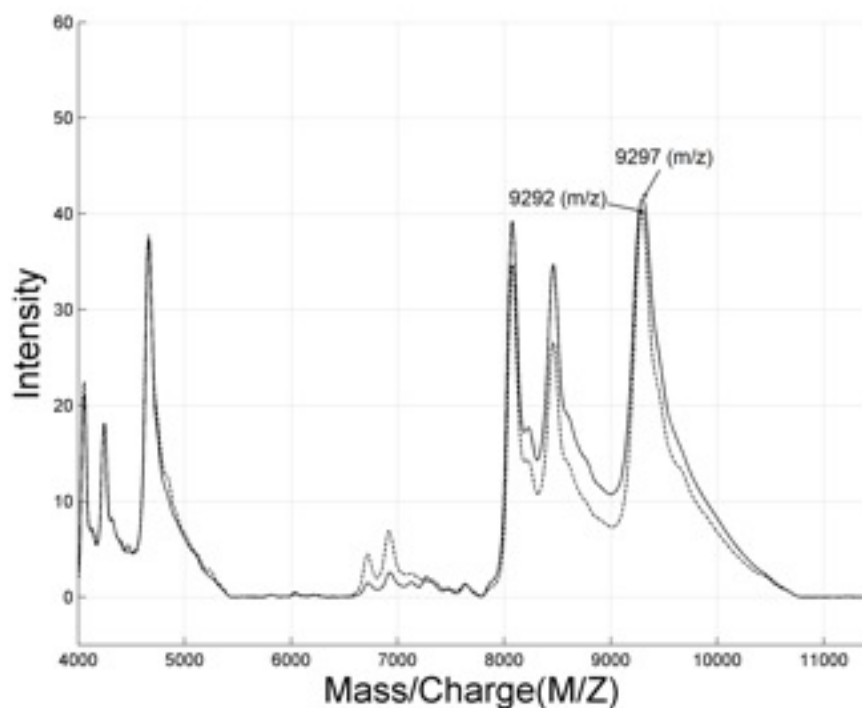


Selected peaks (circles), as resulted for 3 equidistant intensity zones



Selected peaks (circles), as resulted for 5 equidistant intensity zones

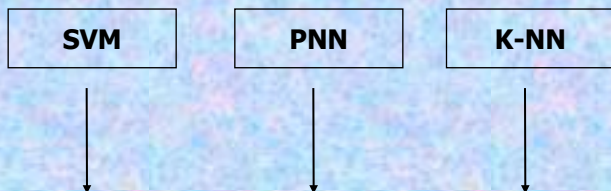
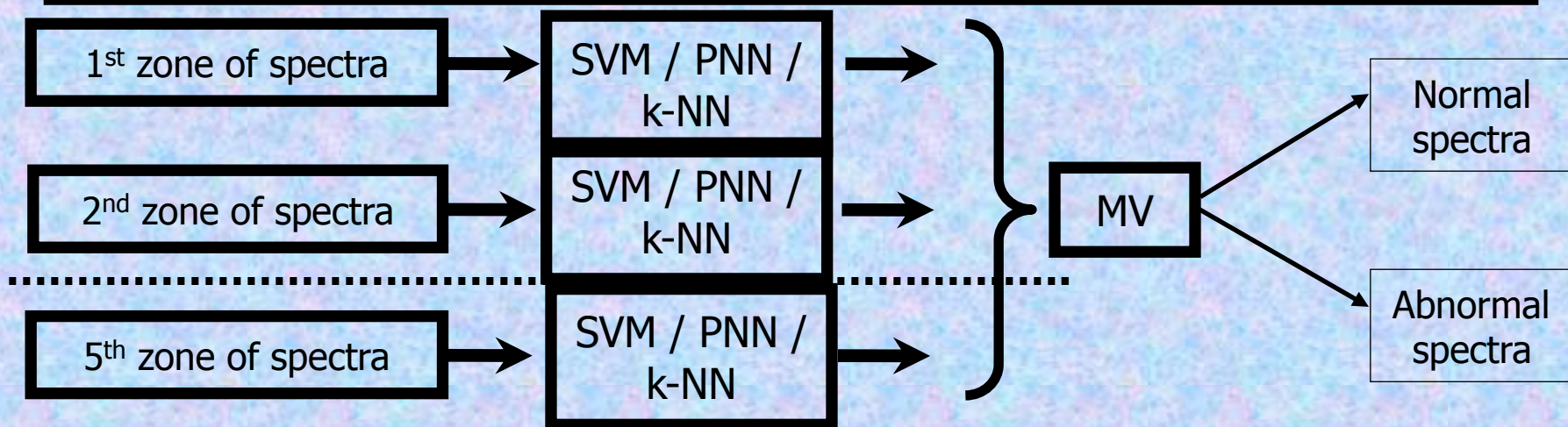
## METHODS (V): Peak alignment



- Peaks of each intensity zone, at each class were initially aligned along the m/z locations to alleviate shifts due to chemical and electronic noise [Baggerly K. A. Bioinformatics,20, 2004]
- Peaks belonging to the same peak-intensity zone across all spectra of the training data set were coalesced, if they differed, with respect to their m/z value, by less than 0.3% relative mass

A region of two spectra demonstrating the “peak-shifting” problem with respect to the x-axis (m/z values).

## METHODS (VI): Classification



Discrimination of normal from abnormal spectra

determination of potential meaningful cancer biomarkers (m/z values), extracted from different MS-spectra intensity regions

- Selection of a single classifier design for each group
- Combined on a single ensemble scheme using a majority vote rule.

## METHODS (VII)

### Evaluation:

- In order to assess the performance of the proposed method and to provide a nearly unbiased estimate of the prediction error rate, the External Cross Validation (ECV) method was utilized [Ambroise C. 2002]
- MS-spectra dataset was randomly separated ten times into 2 subsets: a training dataset (70% of the data) and a testing dataset (30% of the data)
- The training dataset was used for designing an optimum classification scheme
- The testing dataset was used for assessing its predictive performance on unseen MS-spectra.

### Material:

- **MS-spectra from the ovarian-cancer dataset and the prostate-cancer dataset from the National Cancer Institute Clinical Proteomics Database**
- 253 spectra including cases with no evidence of disease and benign cases, and 69 prostate cancer spectra
- 91 controls and 162 ovarian cancer cases

**Furthermore, in an effort to avoid possible sources of distortion of the 'lower' part of MS-spectra by energy-absorbing-molecules, m/z values greater than 1500 in proteomic spectra were also considered for classification**

## RESULTS II: Prostate dataset

Zones	m/z values
1st	1515.2, 1687.9, 4669.0, 6703.7, 6908.9, 8079.5
2nd	3473.3, 3697.1, 4480.3, 9284.5
3rd	2528.2, 2558.8, 2591.0, 2611.7, 2708.0, 3372.0, 3846.4, 4794.5, 5337.0, 6108.2, 9739.5, 14083.7, 18770.5
4th	1637.3, 1746.1, 1771.8, 1787.3, 3864.4, 4929.9, 4987.7, 5113.1, 8465.7, 9121.9
5th	1521.6, 1531.3, 1989.2, 2000.0, 2934.7, 3013.4, 4397.0, 4508.2, 5251.1, 6378.8, 7327.0, 14162.4

Most frequently appearing m/z values in each zone in the 10 runs of the ECV method

## RESULTS II: Prostate dataset

	Proposed Biomarkers	Accuracy (%)	(m/z) > 1500
Petricoin, E. F. 2002	√	90.9	-
Jong, K. 2004.	√	90.7	-
Jong, K. 2004	√	91.3	-
Bhanot, G. 2006	√	91.5	-
Our Method	√	<b>92.5</b>	-
Our Method (>1500)	√	<b>86.7</b>	√

Zones	m/z values
1st	1515.2, 1687.9, 4669.0, 6703.7, 6908.9, 8079.5
2nd	3473.3, 3697.1, 4480.3, 9284.5
3rd	2528.2, 2558.8, 2591.0, 2611.7, 2708.0, 3372.0, 3846.4, 4794.5, 5337.0, 6108.2, 9739.5, 14083.7, 18770.5
4th	1637.3, 1746.1, 1771.8, 1787.3, 3864.4, 4929.9, 4987.7, 5113.1, 8465.7, 9121.9
5th	1521.6, 1531.3, 1989.2, 2000.0, 2934.7, 3013.4, 4397.0, 4508.2, 5251.1, 6378.8, 7327.0, 14162.4

Most frequently appearing m/z values in each zone in the 10 runs of the ECV method

**RESULTS II: Ovarian dataset**

Most frequently appearing m/z values in each zone in the 10 runs of the ECV method

<b>Zone 1</b>		<b>Zone 2</b>		<b>Zone 3</b>		<b>Zone 4</b>		<b>Zone 5</b>	
1671.3	4596.3	1617.5	3532.4	2199.7	5034.4	3203.7	12132.3	10946.2	13162.7
1789.3	5274.5	1657.2	3648.2	2277.4	5276.5	3247.2	13921.7	11035.7	13191.6
1961.7	5958.9	1670.9	3673.0	2437.3	5771.3	5523.4	14152.5	11218.8	13320.5
2080.5	7054.7	1962.2	4266.5	2796.6	6602.9	9857.7	15711.2	11506.3	13937.2
2307.8	7250.9	2080.9	4875.0	2986.0	6655.3	10020.0	17130.4	11593.6	15710.6
2666.4	7379.0	2307.3	5039.9	3161.6	7086.9	10040.6	18478.8	11628.3	
3070.5	7535.7	2543.0	5275.8	3203.7	17101.1	10061.2		12133.8	
3532.7	7965.8	2559.0	6039.8	3673.6		10126.8		12722.1	
3642.5	9168.3	2666.4	6812.7	3745.8		11043.6		12749.5	
3993.6	9436.5	2795.6	6847.8	3822.7		11385.5		12893.2	
4186.4	10525.1	3070.6	7244.9	4264.0		11624.7			
		7383.8	9531.0						

## RESULTS II: Ovarian dataset

Most frequently appearing m/z values in each zone in the 10 runs of the ECV method

Zone 1		Zone 2		Zone 3		Zone 4		Zone 5	
1671.3	4596.3	1617.5	3532.4	2199.7	5034.4	3203.7	12132.3	10946.2	13162.7
1789.3	5274.5	1657.2	3648.2	2277.4	5276.5	3247.2	13921.7	11035.7	13191.6
1961.7	5958.9	1670.9	3673.0	2437.3	5771.3	5523.4	14152.5	11218.8	13320.5
2080.5	7054.7	1962.2	4266.5	2796.6	6602.9	9857.7	15711.2	11506.3	13937.2
2307.8	7250.9	2080.9	4875.0	2986.0	6655.3	10020.0	17130.4	11593.6	15710.6
2666.4	7379.0	2307.3	5039.9	3161.6	7086.9	10040.6	18478.8	11628.3	
3070.5	7535.7	2543.0	5275.8	3203.7	17101.1	10061.2		12133.8	
3532.7	7965.8	2559.0	6039.8	3673.6		10126.8		12722.1	
3642.5	9168.3	2666.4	6812.7	3745.8		11043.6		12749.5	
3993.6	9436.5	2795.6	6847.8	3822.7		11385.5		12893.2	
4186.4	10525.1	3070.6	7244.9	4264.0		11624.7			
		7383.8	9531.0						

	Proposed Biomarkers	Accuracy (%)	(m/z) > 1500
Petricoin, E. F.2002	√	96	√
Sorace, J. M. 2003	√	93.7	√
Alexe, G. 2004	√	92.5	√
Fushiki, T. 2006	√	94	-
<b>Our Method</b>	√	<b>97.18</b>	√

## CONCLUSIONS

1. The methodology of splitting proteomic spectra into intensity regions proved efficient in boosting up the performance of the system under the Majority vote combination rule. Highest accuracy, as compared to other studies, was achieved.
2. The process of splitting the proteomic spectra facilitated the investigation of the peaks according to their intensity. It was found that peaks realizing high intensity values were the list significant in the classification process of both datasets.
3. The proposed peak-intensity zones grouping method facilitated in reducing the number of probable biomarkers, as well as, in assisting researchers by focusing on specific intensity levels for biomarker discovery
4. Proposed method might be of value in the discrimination of normal from cancerous MS-spectra.