# Pattern Recognition System Development for the Classification of Prostate Mass Spectrometry Data

Spyridon Kostopoulos[1], Dimitrios Glotsos[1], Konstantinos Sidiropoulos[2], Panteleimon Asvestas[1], Evripidis Mikos[1], George Sakellaropoulos[3], and Ioannis Kalatzis[1*]

[1]Department of Biomedical Engineering, Technological Education Institute of Athens, Greece, *ikalatzis@teiath.gr

[2]School of Engineering and Design, Brunel University West London, Uxbridge, Middlesex, UB8 3PH, UK

[3]Medical Image Processing and Analysis Group, Laboratory of Medical Physics, School of Medicine, University of Patras, 265 00 Rio, Greece

**Keywords:** Mass spectrometry, prostate cancer, pattern recognition, signal processing

## Abstract

Prostate mass spectrometry (MS) data analysis could contribute to early prostate cancer diagnosis. A pattern recognition system was developed to classify MS data in normal/benign and malignant prostate cases. The system comprises the probabilistic neural network classifier combined with suboptimal feature selection techniques and two different evaluation methods. The external cross validation method was used to evaluate the performance of the integrated system. The optimal feature combinations were achieved with sequential forward floating selection technique combined with the leave-one-out method.

## Introduction

Prostate cancer is the second leading cause of cancer deaths in United States and Canada. The most widely used method for prostate cancer detection is the measurement of the prostate specific antigen (PSA). The PSA diagnostic test exhibits high sensitivity. However, its low specificity confines its use as an early detection biomarker. Mass Spectrometry (MS) data analysis helps on understanding the correlation between proteins/peptides and various diseases as well as the early cancer diagnosis.

In order to contribute towards the MS biosignals analysis, a pattern recognition system was developed for the classification between normal/benign and cancerous prostate MS spectra. This study describes the procedure followed to evaluate an integrated pattern recognition system. Various feature combinations selection methods as well as different evaluation criteria were tested, in order to investigate and compare the capabilities and advantages of each method in obtaining the optimal classification parameters.

## Material and Methods

Data were retrieved from the National Cancer Institute (USA) Clinical Proteomics Database (Petricoin, 2002). Two different classes were built as input to the pattern recognition system. Class 1 (HL-BE) included the feature vectors characterizing 63 normal cases with prostate specific antigen (PSA) <1 (HL) and 190 benign cases with PSA>4 (BE). Class 2 (ML) included feature vectors characterizing 26 malignant cases with PSA from 4 to 10 and 43 cases with PSA>10. Mass spectrometry (MS) signals were preprocessed with resampling, base line

correction, normalization and smoothing with background noise reduction, followed by peak detection and spectra alignment (Roy, 2010). As a result, each MS spectrum was characterized by 170 m/z intervals (Wasinger, 2013), comprising the feature vectors of the classification system.

**Classification.** The classification system included the Probabilistic Neural Network (PNN) classifier (Specht, 1990), equipped with Gaussian activation function and discriminant function given in Eq. 1:

$$f_{\mathrm{C}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}\sigma^{d}} \frac{1}{n} \sum_{i=1}^{n} \exp\left( -\frac{(\mathbf{x}-\mathbf{x}_{Ci})^{T}(\mathbf{x}-\mathbf{x}_{Ci})}{2\sigma^{2}} \right) \quad (1)$$

where $n$ is the number of training patterns, $\mathbf{x}_{Ci}$ is the $i$-th training pattern of class $C$, $d$ is the pattern space dimensionality and $\sigma$ a smoothing factor.
The smoothing factor value was set equal to 0.25 during the training phase and equal to 0.75 during the test phase.

**System evaluation.** The system was evaluated using the External Cross Validation (ECV) method (Ambroise, 2002). Accordingly, each class is divided into two groups. The training group is used exclusively for system training, i.e. to obtain the appropriate parameters that will be used for classification. The test group is used exclusively for system evaluation based on the parameters selected at the previous stage. The test group patterns are formatted as to consist of the optimal features selected in the training phase, while their normalization is performed using the parameters (mean value, standard deviation) of the training group. In this study, class patterns were randomly divided into two groups with 2:1 ratio; the training group comprised the 2/3 of the total number of cases, while the rest 1/3 consisted the test group. The whole procedure was repeated 10 times, each with different random data division, in order to minimize bias. The mean overall accuracy of the test group classification, over the 10 repetitions, is considered the expected accuracy of the system when generalized on unknown data.

**Training phase.** During training, optimal feature vectors combinations were tested using the exhaustive search (EXS) and the sequential forward selection (SFS) techniques. In particular, an exhaustive search was performed with maximum feature vector dimensionality equal to 2 followed by SFS with maximum feature vector dimensionality equal to 6 (EXS-SFS). Moreover, sequential forward floating selection (SFFS) was also used. Each feature combination performance was evaluated using resubstitution (RESUB) and leave-one-out (LOO) methods. The overall accuracy (OA) of a classifier and the J3 class separability criterion were used for the optimal feature combination selection (Theodoridis, 2003). The classifier used at this phase was PNN equipped with Gaussian activation function and smoothing factor value equal to 0.25.

**Test phase.** Test group patterns were classified based on the feature combination selected from the previous stage. The optimal feature combination is characterized by the highest value of the criterion used (OA, J3) with the lowest feature vector dimensionality. PNN classifier, with the Gaussian activation function and smoothing factor value equal to 0.75, was used to classify the test group. The overall accuracy was calculated from the corresponding confusion matrix.

## Results and Discussion

The test group classification accuracies of the two MS spectra classes, for each evaluation method and for each feature selection technique, with the corresponding criteria values calculated at the training phase, are presented in Table 1.

| Accuracies after training using various techniques & evaluation criteria | | | PNN test phase accuracies | | |
|---|---|---|---|---|---|
| Feature selection technique | Evaluation criterion | Critetrion value | Overall Accuracy (%) | Specificity (%) | Sensitivity (%) |
| EXS-SFS | J3 | 1.17 | 78.9 | 78.2 | 81.3 |
| SFFS | J3 | 1.16 | 78.4 | 76.0 | 87.4 |
| EXS-SFS | PNN-RESUB | 100% | 69.8 | 75.5 | 48.7 |
| SFFS | PNN-RESUB | 100% | 66.6 | 72.0 | 46.5 |
| EXS-SFS | PNN-LOO | 89% | 75.5 | 76.8 | 70.4 |
| SFFS | PNN-LOO | 97% | 80.8 | 84.4 | 67.4 |

Table 1. Mean values and standard deviations of the test group overall accuracies, using the PNN classifier with the optimal feature combination, for 10 repetitions of ECV.

The large number of features (m/z values) used in this study, which is native in MS data, necessitates the use of suboptimal feature selection techniques (Theodoridis, 2003). The SFFS technique proved computationally fast in the specific dataset, resulting in optimal feature combinations of high dimensionality, something difficult to achieve using EXS-SFS or plain exhaustive search. The SFFS technique gave the highest mean overall accuracy using the robust leave-one-out method during both training and (most importantly) test phase.

The highest overall accuracy of the test group (80.8%) was achieved with the PNN classifier, using the LOO method and SFFS technique. Feature vectors retrieved at each repetition of the ECV procedure comprised of a large number of features (11 to 31, mean = 20).

This procedure (OA/PNN with LOO, feature vectors search with SFFS) also resulted to the highest specificity value (84.4%). This is considered very important, given the low specificity value of the PSA test on prostate cancer detection (Pannek, 1998). Nevertheless, the highest sensitivity was achieved using the J3 criterion (87.4% with SFFS), with a comparatively low number of features (1 to 6, mean = 4). This could probably lead to the conclusion that the data structure is relatively simple, at least regarding benign from malignant class separability. This is further supported by the simple statistics of the J3 criterion (larger dispersion between than within classes).

## Concluding Remarks

A pattern recognition system was developed for the classification of prostate cancer MS spectra. The optimal feature combinations were achieved with the SFFS technique combined with the LOO method and the PNN classifier. This pattern recognition scheme resulted in the highest overall accuracy as well as the highest specificity. Highest sensitivity though, was attained using the J3 criterion, with a much lower feature vector dimensionality.

## Acknowledgements

This research has been co-funded by the European Union (European Social Fund) and Greek national resources under the framework of the "Archimedes III: Funding of Research Groups in TEI of Athens" project of the "Education & Lifelong Learning" Operational Programme.

## References

Ambroise, C., and McLachlan, G., 2002, Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences of the USA, 99(10), 6562–6566.

Pannek, J., and Partin, A. W., The role of PSA and percent free PSA for staging and prognosis prediction in clinically localized prostate cancer. Seminars in Urologic Oncology, 16(3), 100-105.

Petricoin Iii, E. F., Ornstein, D. K., Paweletz, C. P., Ardekani, A., Hackett, P. S., Hitt, B. A., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C. B., Levine, P. J., Linehan, W. M., Emmert-Buck, M. R., Steinberg, S. M., Kohn, E. C., and Liotta, L. A., 2002, Serum proteomic patterns for detection of prostate cancer. Journal of the National Cancer Institute, 94, 1576-1578.

Roy, P., Truntzer, C., Maucort-Boulch, D., Jouve, T., and Molinari, N., 2011, Protein mass spectra data analysis for clinical biomarker discovery: A global review. Briefings in Bioinformatics, 12(2), 176-186.

Specht, D.F., 1990, Probabilistic Neural Networks. Neural Networks, 3, 109-118.

Theodoridis, S., and Koutroumbas, K., 2003, System evaluation (London, U.K.: Academic Press, 2nd ed.).

Wasinger, V. C., Zeng, M., and Yau, Y., 2013, Current status and advances in quantitative proteomic mass spectrometry. International journal of proteomics, article ID 180605, 12 pages.