International Journal of Innovative Research in Computer and CommunicationEngineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

Artificial Neural Network for Prostate Cancer Classification from Mass Spectrometry Data

Masoud Alajmi

Assistant Professor, Computer Engineering, Taif University, Taif, Kingdom of Saudi Arabia

ABSTRACT: Mass spectrometry (MS) is a technique for molecular analysis. MS is widely used for detecting tumors and monitoring progression. In this work, MS data was used for prostate cancer classification. Using MS for classification will be a very computational process due to three main challenges. MS spectra usually have high dimension and are contaminated with noise. In addition, most available data sets are considerably small and unbalanced, which leads to overfitting. Many methods have been proposed to address the dimensionality and noise, but not the number of samples. We propose a simple and effective augmentation method to artificially add more samples to the data set. Then, we propose Artificial Neural Network (ANN) as a classifier. The results were compared with state-of-the-art methods. The proposed method outperformed all compared techniques.

KEYWORDS*:* Mass Spectrometry, ANN, Classification

I. INTRODUCTION

a. Mass Spectrometry (MS)

MS is a method of analysis that can generate ions produced by both inorganic and organic molecules. The technique measures the mass-to-charge ratio (m/z) of the ions and identifies them, both qualitatively and quantitatively, by way of their respective m/z ratio intensities. Liquid chromatography can be coupled with mass spectrometry (LC/MS) in order to separate ions before they are detected. The procedural components of mass spectrometry are (1) ionization, (2) separation, and (3) detection of ions in a gas phase. Using ionization source, sample molecules are first ionized, then an LC-column allows sample components to be separated. While a protein digest is separating, peptides elute from the column at various times, which are recorded as retention times (rt). After analysis of the peptides using a mass-to-charge ratio, the relative intensities of m/z signals are recorded by the mass analyzer. The detector then measures the ions emerging from the mass analyzer [1] [2] [3].

b. Artificial Neural Network (ANN)

Artificial neural networks, in the form of either mathematical tools or physical devices, are designed to function like biological neural systems. Their component parts, referred to as “artificial neurons,” act as building blocks that are similar to the structure of actual biological neurons. Every biological neuron has three primary components: dendrites, soma, and axon.Similarly, an artificial neuron has three major parts: inputs (or “dendrites”), transformation function (“soma”), and output (“axon”). The structures correspond so closely that the labels used for biological neurons are frequently also utilized for parts of artificial neurons. Modern neural networks utilize data analysis and non-linear statistical techniques to describe complicated connections between inputs and outputs among them Bayesian inference methods, graph theory, and geometry. Researchers are increasingly turning to Bayesian inference methods (named for Thomas Bayes). Graph theory and geometry are effective for mapping neural networks, evaluating their capabilities, and determining pattern classification. There are many kinds of artificial neural networks, each designed to find the best solution for a specific problem [4].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

II. RELATED WORK

Generally, the high dimensionality combined with noise peaks of MS spectra are the main challenges for classifications. Many methods have been proposed for MS classification: Principal component analysis with discriminant function analysis [5] [6], decision tree learning [7] and Linear discriminant analysis (LDA) [8]. Other methods focused on identifying biomarker peaks in MS samples to distinguish between cancer categories. Sophistically statistical methods were applied to identify biomarker peaks [9] [10]. There is another issue related to MS classification: the number of samples in data sets is small, and the classifier performance is affected. To the best of our knowledge, no approach has been proposed to address the data set size. In this work, we applied ANN for MS classification and proposed simple and effective augmentation method to improve the classification performance.

III. MATERIAL AND METHODS

a. Data set structure

In this work, we use data sets from the NIH and FDA [11]. The data can be considered as a matrix with dimension 15200x338, where 15200 represents the range of m/z value and 338 is the total number of samples. The data set contains four unbalanced classes that are listed based on Prostate Specific Antigen (PSA) level. Table 1 shows this data listed according to a Prostate Specific Antigen (PSA) level. 62 samples showedno evidence of disease; 209 were benign; 25 showed presence of prostate cancer with a PSA level of 4 – 10, and 42 showed prostate cancer with a PSA level of > 10ng/mL. Fig.1 shows a sample from each category.

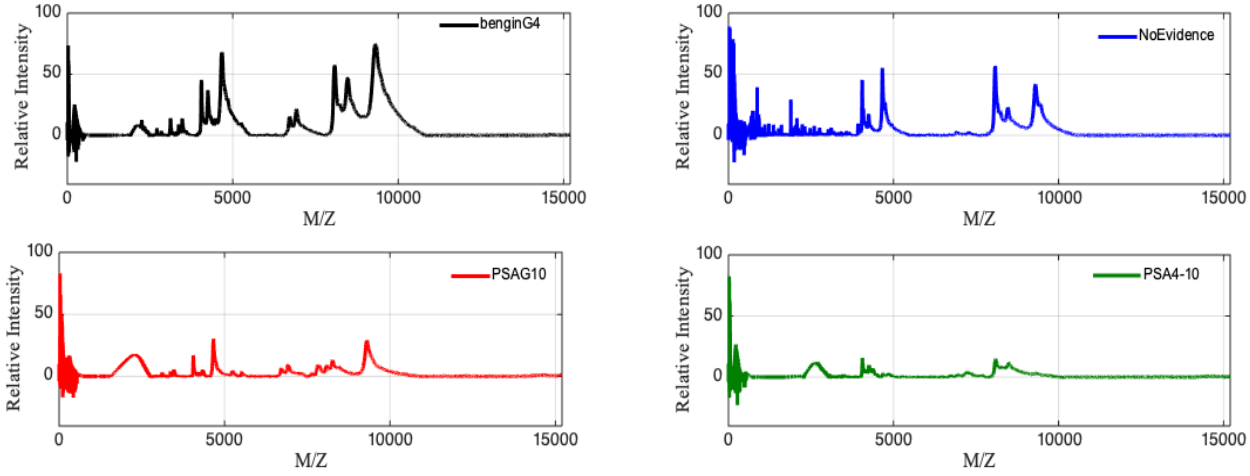

Fig.1: Sample from each category in prostate cancer data set.

b. Preprocessing data:

Due to a high variation of amplitude peaks, the normalization process was used to set the intensities into new values in range with [0,1]. All the samples werenormalized based on the following formula:

$$y_j^{normal} = \frac{y_j - \min(y_j)}{\max(y_j) - \min(y_j)} \qquad (1)$$

where $max(y_j)$and $min(y_j)$ are the maximum and minimum intensity peaks respectively, for j^{th}sample. In addition, the sample range (15200) can be filtered based on the feature's values. All features in the high range were zeros, and all samples were filtered to be 10300 features.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

c. Data augmentation

Neural network performance can be significantly improved by using more combinations of training sets. A small number of samples in a data set leads to overfitting. Data augmentation alleviates this by using existing data more effectively. In this paper, we propose a data augmentation approach for MS data. Basically, a new sample in a particular category can be generated using the average of two other samples belonging to the same category,

$$y_{j,i}^{Aug} = \frac{y_{j,i+1}+y_{j,i+2}}{2} \quad (2)$$

where *j* is the class label, and *i*is a sample in the same class.

The augmented method can be applied to every class sample in a training set. If the number of samples in a category is n, the maximum number of generated samples will be $n(n-1)/2$. The advantage of this is that the training set can be balanced by dynamically choosing the augmented samples based on the category. Fig.2 shows an example of the augmented sample in a specific range. To get more variety, we include the correlation as the parameter to find the samples that we needed to combine. Two samples should be combined based on the lowest correlation coefficient.

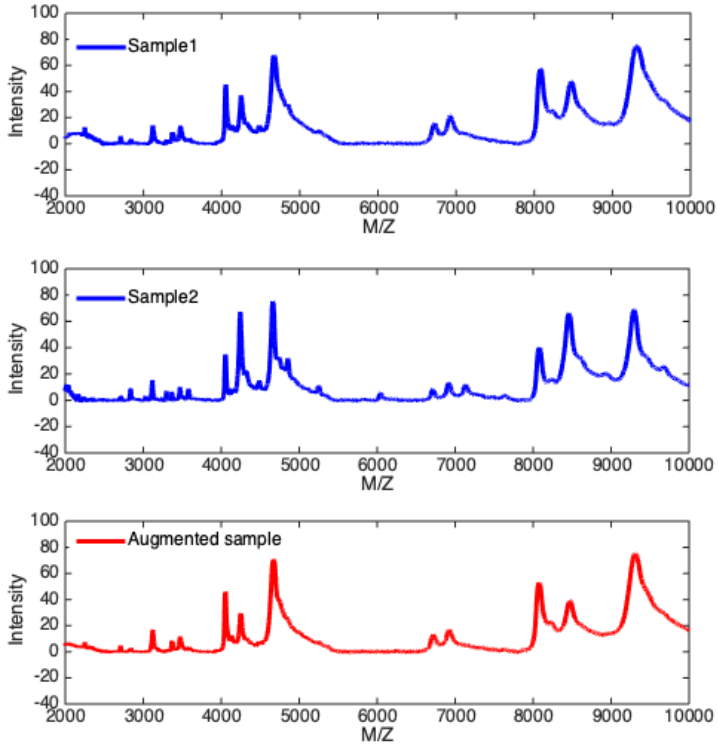

Fig.2: An example of augmentation two samples from the same category.

d. Data classification

The data extracted from prepressing acts as an input to the neural network, as shown in Fig. 3. The designed neural network has two hidden layers. The feed-forward back-propagation method was used to select for the learning rule. The cancer classes were represented in the output layer. The activation functions for hidden layer neurons were rectified linear unit functions, and output layer was the normalized exponential function. In addition, performance of the network was calculated according to the accuracy.

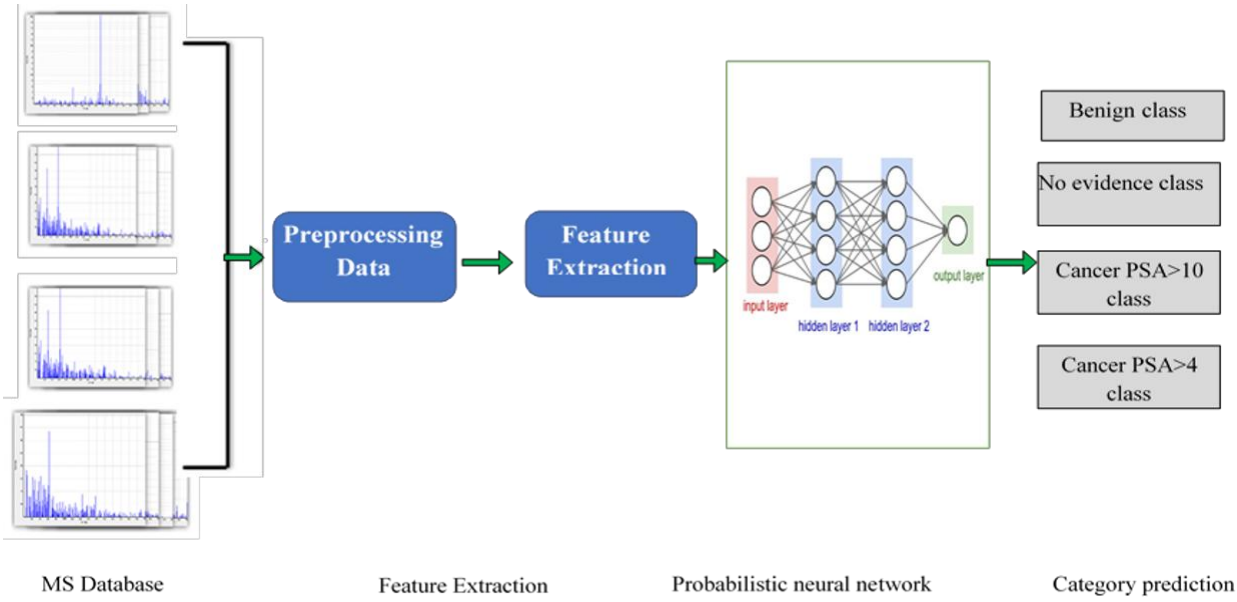

Fig.3: General structure for the proposed method.

IV. RESULTS AND DISCUSSIONS

Determining the ANN architecture is the most important component and is determined by a trial and error process. The number of neurons used for all layers was 512, the maximum number of iterations(epochs) was selected to be 100. A K-fold cross validation (k=10) is carried out for all data sets. To observe the preprocessing and augmentation steps, the data set was applied after each step according to different percentages of each data set. Data was augmented by 10% of total samples. Fig. 4 shows the results of accuracies and loss functions. A basic observation is that the accuracy and loss function were improved after every preprocessing steps.

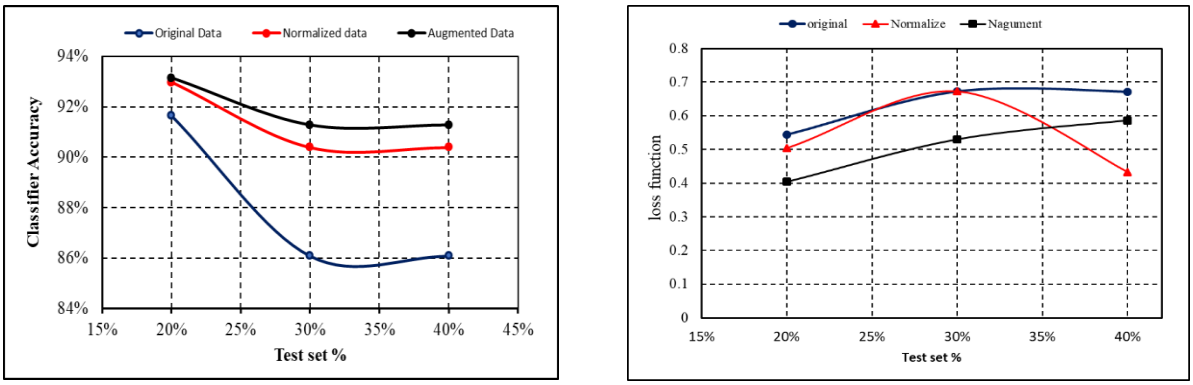

Fig.4: Classifier performance comparison for prepressing steps.

Table 1: Classifiers performance comparison

Test set (%)	Classifier			
	KNN	SVM	Dtree	ANN
10	95.0 ± 0.02	97.50 ± 0.018	91.18± .03	98.81± 0.32
20	94.0 ± 0.02	97.32 ± 0.010	89.29± 0.02	97.51± 0.48
30	93.0 ± 0.01	97.01± 0.007	88.70± 0.01	97.61± 0.44
40	95.0 ± 0.05	96.60 ± 0.012	88.30±0 .02	97.32± 0.13
50	91.0 ± 0.01	95.84 ± 0.008	86.85± 0.01	96.17± 0.26

The performance obtainedfrom our approach for the MS dataset were compared with the results obtained from three common classifier methods: support vector machine (SVM) [12], k-nearest-neighbors (KNN) [13], and decision tree [7]. In all methods, the results were based on 10- cross validation. The input data to all classifiers were preprocessed and augmented. The number of augmented samples were dynamically changed based on the number of training samples in each class. Table 1 summarizes the average training and test performances of the proposed approach and compared with other methods.

Fig.5 displays the accuracy comparison of the four classification methods, including our approach. The results reveal that the ANN with data augmentation demonstrates outstanding performance, even with a small number of training samples.

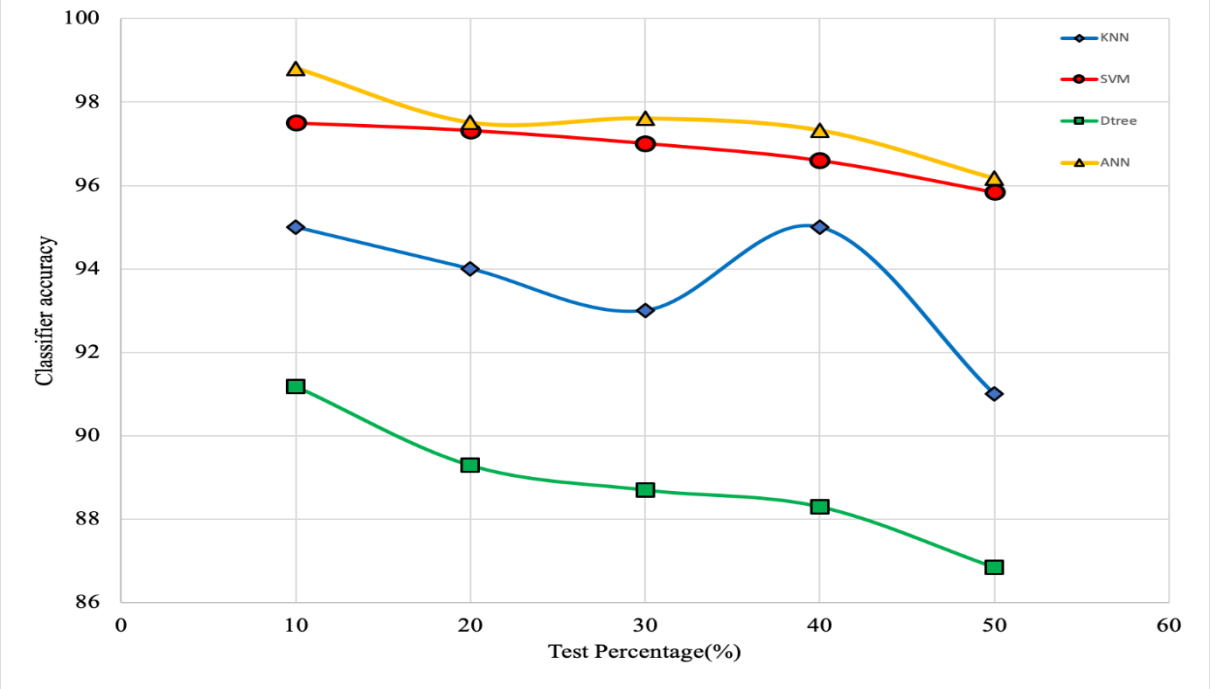

Fig.5: The classifier's accuracy vs. percentage of test set for all methods.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

V. CONCLUSION AND FUTURE WORK

In this work, ANN and a simple augmentation method are proposed to address the challenges of MS classification. The proposed approach outperforms other classification methods. The data was augmented dynamically based on the samples in each category. In the future, the ANN network parameters need to be investigated in order to optimize the classifier performance.

REFERENCES

[1] E. Hoffman, "Tandem mass spectrometry: a primer," *Journal of mass spectrometry*, vol. 31, no. 2, pp. 129-137, 1996.

[2] K. Awedat, M. Alajmi and J. R. Springstead, "Compressive Sensing as a Method for Spectrometry Reconstruction," in *2017 IEEE International Conference on Electro Information Technology (EIT)*, Lincoln, NE, USA, 2017.

[3] K. Podwojski, A. Fritsch, D. C. Chamrad, W. Paul, B. Sitek, K. Stühler, P. Mutzel, C. Stephan, H. E. Meyer, W. Urfer and others, "Retention time alignment algorithms for LC/MS data must consider non-linear shifts," *Bioinformatics*, vol. 25, no. 6, pp. 758-764, 2009.

[4] V. E. Bondarenko, "Artificial Neural Networks.," 2019. [Online]. Available: http://search.ebscohost.com.sdl.idm.oclc.org/login.aspx?direct=true&db=ers&AN=94981752&site=eds-live. [Accessed 2020].

[5] M.J.Baker,M.D.Brown,E.Gazi,N.W.Clarke,J.C.Vickerman,andN.P.Lockyer, "Discrimination of Prostate Cancer Cells and Non-malignant Cells Using Secondary Ion Mass Spectrometry," *Journal of Educational Psychology*, vol. 133, pp. 175-179, 2008.

[6] W. J. Krzanowski, "Principles of Multivariate Analysis: a User's Perspective," Oxford University Press, Revised Ed., 1988.

[7] Song, Yan-Yan and Ying, LU, "Decision tree methods: applications for classification and prediction," *Shanghai archives of psychiatry*, vol. 27, p. 130, 2015.

[8] Baolin Wu, Tom Abbott, David Fishman, Walter McMurray, Gil Mor, Kathryn Stone, David Ward, Kenneth Williams and Hongyu Zhao, "Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data," *Bioinformatics*, vol. 19, pp. 1636-1643, 2003.

[9] Wang, David L and Xiao, Chuanguang and Fu, Guofeng and Wang, Xing and Li, Liang}, "Identification of potential serum biomarkers for breast cancer using a functional proteomics technology," *Biomarker research*, vol. 5, p. 11, 2017.

[10] hen, Yingrong and Ma, Zhihong and Min, Lishan and Li, Hongwei and Wang, Bin and Zhong, Jing and Dai, Licheng, "Biomarker identification and pathway analysis by serum metabolomics of lung cancer," *BioMed research international*, vol. 2015, 2015.

[11] E. F. Petricoin III; D. K. Ornstein, C. P. Paweletz, A. Ardekani, P. S. Hackett, B. A. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood et al., "Serum pro- teomic patterns for detection of prostate cancer,," *Journal of the National Cancer Institute*, pp. 1576-1578, 2002.

[12] Bottou, Léon, and Chih-Jen Lin, "Support vector machine solvers," *Large scale kernel machines*, vol. 3, pp. 301-320, 2007.

[13] Lüschow, Andreas, and Christian Wartena., "Classifying medical literature using k-nearest-neighbours algorithm.," 2017.