Premalignant Pancreatic Cancer Diagnosis Using Proteomic Pattern Analysis

Zaw Zaw Htike and Shoon Lei Win
Faculty of Engineering, IIUM, Kuala Lumpur, Malaysia
Email: zaw@ieee.org

Abstract—**Pancreatic cancer is one of the deadliest cancers due to the fact that it does not exhibit symptoms in the early stages. Furthermore, when pancreatic cancer gets diagnosed, it is usually too late. Consequently, early diagnosis is highly essential. The dawn of proteomics has brought with it a glimpse of hope of uncovering biomarkers that can be indicative of early pancreatic cancer. Proteome profiling techniques have become popular in the recent years to try to make sense of high-dimensional proteomic data and to find discrepancies between proteomes of healthy samples and cancerous samples. However, the high dimensionality of proteomics data coupled with small sample size poses a challenge. In this paper, we propose a framework using a hybrid logistic tree technique together with a feature selection technique to diagnose premalignant pancreatic cancer. We have validated our framework on a pancreatic cancer peptide mass spectrometry dataset. Satisfactory preliminary experimental results demonstrate the efficacy of our framework.**

Index Terms—**pancreatic cancer, proteomic analysis, pattern recognition**

I. Background

Pancreatic cancer is considered to be among the notorious cancers with extremely high mortality rate. It remains as one of the major unsolved healthy problems today [1]. In fact, only about 4% of the patients survive 5 years or longer after being diagnosed [2]. The rest of the patients who have been diagnosed with pancreatic cancer develop metastasis and die [1]. This is because when pancreatic cancer gets diagnosed, it is usually too late. Conventional methods of detecting pancreatic cancer rely solely on skilled physicians with the help of medical imaging, peritoneal cytology, and tumor markers such as serum cancer antigen (CA) 19-9 to detect symptoms which usually appear in late stages of cancer [3]. In fact, conventional imaging methods sometimes fail to detect small lesions in the early stages because of the fact that retroperitoneal anatomical positions obscure imaging diagnosis. Furthermore, late stages of pancreatic cancer exhibit signs of great resistance to anticancer treatment, resulting in poor diagnosis. Therefore, there is an urgent need to develop new techniques to diagnose pancreatic cancer in its very early stages. There are transcriptome-based techniques that detect pancreatic cancer using gene expression analysis [4]-[6]. However, these techniques are invasive as a biopsy of the suspected cancer tissue has to be extracted. Proteomics techniques have a significant advantage because early cancer can potentially be detected from a simple drop of serum. Mass spectrometry is a technique that could detect the presence of thousands of low molecular weight proteins and peptides in a drop of serum in the form of a ‘mass spectrum’.

Proteome profiling techniques have become popular in the recent years to try to make sense of high-dimensional proteomic data in the form of a mass spectrum and to find discrepancies between proteomes of healthy samples and cancerous samples. Ensemble techniques are very popular in proteome analysis. For example, Bhattacharyya *et al.* [7] utilized a two-step multivariate analysis procedure comprising regression trees to distinguish pancreatic cancer serum samples from control serum samples. Li and Ngom [8] proposed a high dimensional linear machine to diagnose pancreatic cancer. Ge *et al.* [2] compared the prediction performances of a single decision tree algorithm C4.5 with six different decision-tree based classifier ensembles. They claimed that ensemble classifiers always outperformed single decision tree classifiers. Existing techniques have not achieved high accuracy rates. Conventional data mining techniques do not perform well because of the high dimensionality of input data and scarcity of training samples. Furthermore, mass spectra are usually corrupted with a great deal of noise due to random errors, systematic errors, and sample contamination. Therefore, a technique that could address all these issues is required. The rest of the paper is organized as follows. Section 2 presents our proposed framework. Section 3 describes the experimental results and Section 5 concludes this paper.

II. Proteomic Data Analaysis

A. Overview

The goal of premalignant pancreatic cancer diagnosis is to predict, given a mass spectrum derived from a serum sample, whether or not the sample comes from a patient with early pancreatic cancer. We propose a three-layered framework that consists of pre-processing, feature selection, and classification as shown in Fig. 1.

B. Preprocessing

A typical mass spectrum contains intensity measurements at thousands of m/z ratios. Two steps are

Manuscript received July 1, 2014; revised September 4, 2014.

doi: 10.12720/jomb.4.4.288-292

performed in pre-processing: base-line correction and smoothing. Base-line correction is necessary because a major of the m/z ratios may have non-zero intensity values or spurious peaks because of systematic error, background noise, and chemical noise. Therefore, the true mass spectrum without the contaminants should be estimated. We propose a 'top-hat' filter to perform base-line correction. It entails subtracting the observed spectrum its morphological opening [9]. Spectrum smoothing is then performed next in order to alleviate very high frequency components. We proposed a wavelet noise removal technique. It entails dividing the mass spectrum into components of different scales and estimating the wavelet coefficients [10]. Coefficients corresponding to high frequency components are then discarded in order to smoothen the spectrum. Fig. 2 illustrates a proteomic mass spectrum before and after wavelet de-noising.

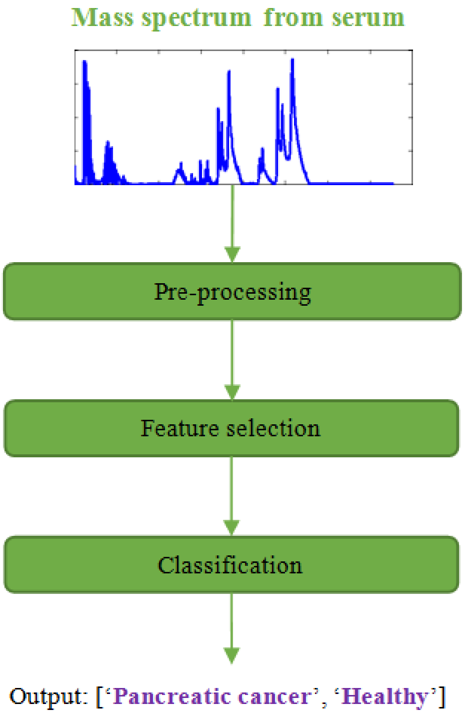

Figure 1. High-level flow diagram of early pancreatic cancer diagnosis system.

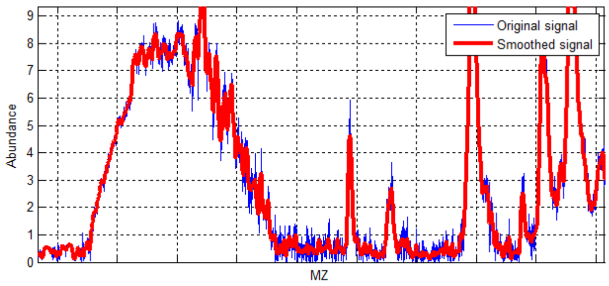

Figure 2. Smoothed mass spectrum after de-nosing.

C. Feature Extraction

A typical proteomic mass spectrum contains tens of thousands of m/z intensity values. The complexity of any machine learning classifier depends upon the dimensionality of the input data [11]-[13]. Generally, the lower the complexity of a classifier, the more robust it is [14]-[17]. Furthermore, not all the intensity values of a mass spectrum might be responsible for cancer diagnosis. Therefore, we need to have a feature selection process. RELIEF [18] is a well-known feature selection algorithm for binary classification. It offers numerous advantages suitable for the problem of cancer diagnosis from proteomic mass spectrum. One of its advantage is that it is highly tolerant to noise and feature interactions. However, it cannot cope with low number of training samples. Therefore, we propose RELIEFF [19] which performs reliable probability estimation, making it capable of coping low number of training samples. We use RELIEFF to select 300 best m/z ratios that best discriminate pancreatic cancer. A proteomic mass spectrum is now represented by a 300-dimensional feature vector.

D. Classification

In the world of data mining and machine learning, there are two popular classes of algorithms: logistic regression- based and tree-based [20]. Each of them has advantages and disadvantages [21]. On the one hand, the former tries to fit simple models to the complex proteomic data, resulting in low variance but potentially high bias. On the other hand, the latter usually utilizes information theoretic metrics such as information gain to build tree-like structures, resulting in low bias but often high variance [20]. Studies have shown that neither of these classes consistently outperform the other and that relative performance depends strongly upon the nature of the dataset [22]. To fuse the best of both worlds, we propose a hybrid technique called a logistic model tree [20] to classify 300-dimensional feature vector. The proposed logistic model tree applies LogitBoost with simple regression functions as base learners in order to fit the logistic models.

III. Experiments

We tested our proposed system using a dataset from the University of Pennsylvania [23]. The dataset contains 181 serum samples where 80 samples are pancreatic intraepithelial neoplasia samples and the remaining 101 samples are healthy or control samples. The mass spectrum of each serum sample contains 6771 m/z ratios that range from 800 to 11992.91.

We carried out a leave-one-out cross-validation where one sample was held out as the validation data while the remaining samples served as training data. The whole process was repeated multiple times such that each sample got held out exactly once as the validation data. The results were then averaged to produce an estimator to the accuracy of the proposed pancreatic cancer diagnosis system. Throughout all the experiments, we used the minimum number of boosting iterations of 50, the maximum number of boosting iterations of 1500, and the heuristic threshold value of 60 as parameters of the logistic model tree. Table I lists the summary of the leave-one-out cross-validation results. The system correctly classified a total of 134 out of 181 samples with an accuracy rate of 74.0331% and an error rate of 25.9669%. Kappa coefficient, which measures inter-rater agreement of predicted values with the true values over all the trials of the leave-one-out cross-validation, was found to be 0.4673. It means that the individual predictions are fairly consistent across multiple trials. MAE and RMSE were found to be 0.3858 and 0.4370 respectively, which were fairly small. RAE and RRSE were found to be significantly large. However, the RAE and RRSE metrics are not very meaningful in the task of classification. Table II displays the detailed results by output class. The true positive (TP) rate of the cancer class is lower than that of the control class. Furthermore, the false positive (FP) rate for the cancer class is also lower than that of the control class. This implies that the system produces more negative predictions than positive predictions. This maybe because of statistical bias caused by having more control samples than cancer samples.

TABLE I. LOOCV RESULTS SUMMARY.

Metric	Value
Correctly classified instances	134 (74.0331 %)
Incorrectly classified instances	47 (25.9669 %)
Kappa coefficient	0.4673
Mean absolute error (MAE)	0.3858
Root mean squared error (RMSE)	0.4370
Relative absolute error (RAE)	77.7790 %
Root relative squared error (RRSE)	87.5146 %

TABLE II. DETAILED RESULTS BY OUTPUT CLASS.

	Cancer	Control
True positive (TP) rate	0.650	0.812
False positive (FP) rate	0.188	0.350
Precision	0.732	0.745
Recall	0.650	0.812
F-score	0.689	0.777
ROC Area	0.784	0.784
Matthews correlation coefficient	0.470	0.470
Precision-recall curve area	0.697	0.793

Fig. 3 illustrates the ROC curve for the cancer class and Fig. 4 illustrates the ROC curve for control class. In summary, the proposed system has managed to produce satisfactory results. The overall accuracy is not that high because the University of Pennsylvania dataset is a very challenging dataset. Fig. 5 illustrates mass spectra of six pancreatic intraepithelial neoplasia samples and six control samples. As shown in the figure, there seems to be no noticeable difference in mass spectra of pancreatic intraepithelial neoplasia sample and six control samples. This shows the difficulty in detecting premalignant pancreatic cancer.

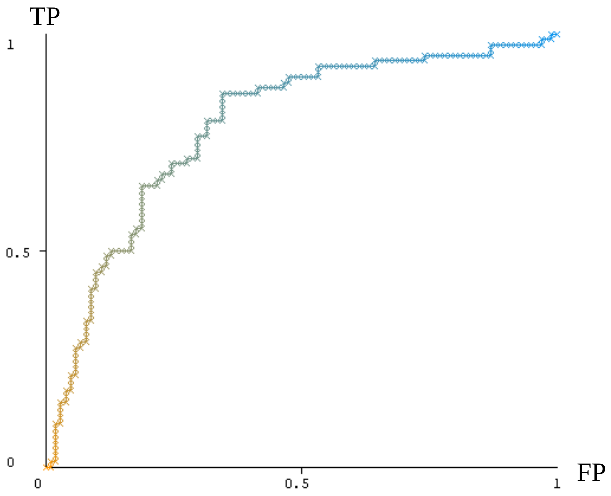

Figure 3. ROC curve for premalignant pancreatic cancer

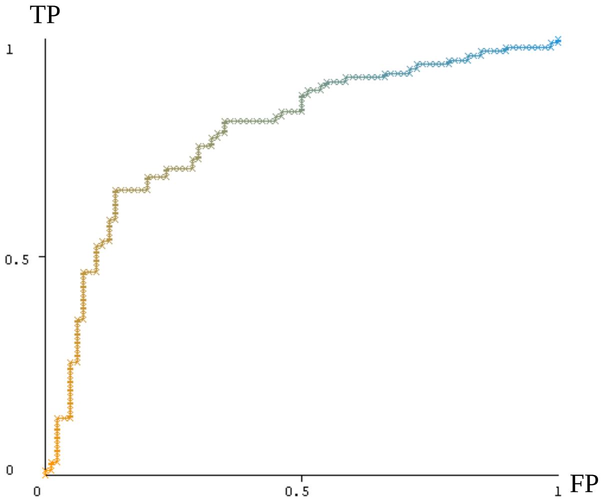

Figure 4. ROC curve for control class

IV. CONCLUSION

We have presented a machine learning based approach to diagnose premalignant pancreatic cancer from serum samples. Given the mass spectrum of a serum sample, the system predicts whether the serum shows signs of premalignant pancreatic cancer. We have carried out experiments on a dataset from the University of Pennsylvania. This proposed system has achieved an accuracy of 74.0331% in early premalignant pancreatic cancer detection for this dataset. The accuracy is not that high because this is a very challenging problem owing to the fact that in the early stages of cancer, there are only miniscule differences in the proteomes. However, the preliminary experimental results are quite promising. As future work, we would like to perform optimization of the system parameters to further boost the performance of the system. We also would like to test this framework on a wide range of other types of cancer.

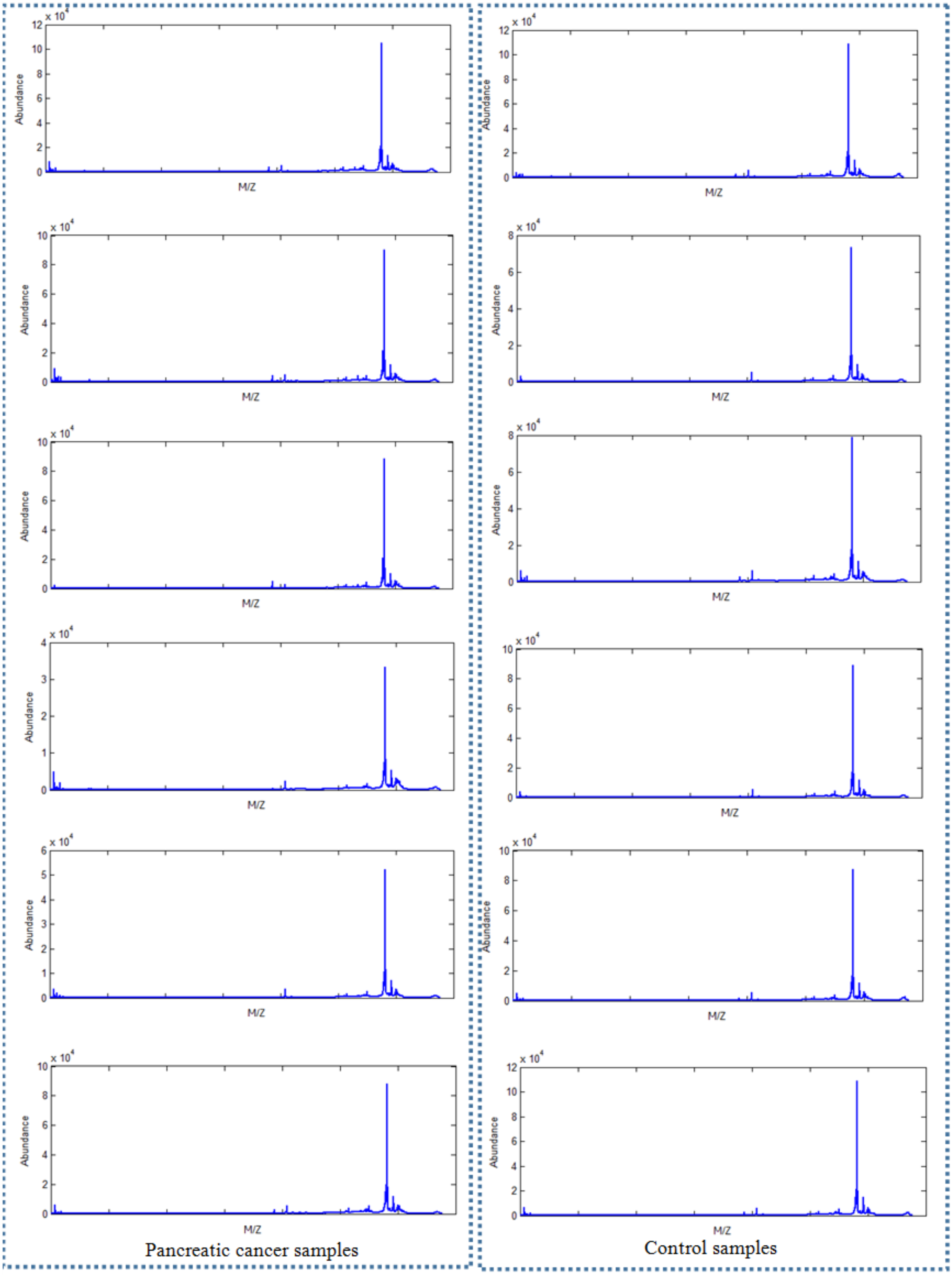

Figure 5. Mass spectra of pancreatic cancer samples (left) vs. mass spectra of normal/control samples (right).

References

[1] K. Li, *et al.*, “Pancreatic cancer,” *The Lancet*, vol. 363, pp. 1049-1057, 2004.

[2] G. Ge and G. W. Wong, “Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles,” *BMC Bioinformatics*, vol. 9, pp. 275, 2008.

[3] A. Ghazale, *et al.*, “Value of serum IgG4 in the diagnosis of autoimmune pancreatitis and in distinguishing it from pancreatic cancer,” *The American Journal of Gastroenterology*, vol. 102, pp. 1646-1653, 2007.

[4] S. L. Win, *et al.*, “Cancer recurrence prediction using machine learning,” *International Journal of Computational Science and Information Technology*, vol. 6, 2014.

[5] S. L. Win, *et al.*, "Gene expression mining for predicting survivability of patients in early stages of lung cancer," *International Journal on Bioinformatics & Biosciences,* vol. 4, 2014.

[6] S. L. Win, *et al.*, "Cancer classification from DNA microarray gene expression data using averaged one-dependence estimators," *International Journal on Cybernetics & Informatics*, vol. 3, 2014.

[7] S. Bhattacharyya, *et al.*, "Diagnosis of pancreatic cancer using serum proteomic profiling," *Neoplasia,* vol. 6, pp. 674-686, 2004.

[8] Y. Li and A. Ngom, "Diagnose the premalignant pancreatic cancer using high dimensional linear machine," presented at the Proceedings of the 7th IAPR international conference on Pattern Recognition in Bioinformatics, Tokyo, Japan, 2012.

[9] K. R. Coombes, *et al.*, "Improved peak detection and quantification of mass spectrometry data acquired from surface‐enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform," *Proteomics,* vol. 5, pp. 4107-4117, 2005.

[10] E. Alpaydin, *Introduction to Machine Learning*, 2nd ed.: The MIT Press, 2010.

[11] K. R. Coombes, *et al.*, "Pre-processing mass spectrometry data," in *Fundamentals of Data Mining in Genomics and Proteomics*, Ed: Springer, 2007, pp. 79-102.

[12] Z. Z. Htike and S. L. Win, "Recognition of promoters in DNA sequences using weightily averaged one-dependence estimators," *Procedia Computer Science,* vol. 23, pp. 60-67, 2013.

[13] Z. Z. Htike and S. L. Win, "Classification of eukaryotic splice-junction genetic sequences using averaged one-dependence estimators with subsumption resolution," *Procedia Computer Science,* vol. 23, pp. 36-43, 2013.

[14] E. E. M. Azhari, *et al.*, "Brain tumor detection and localization in magnetic resonance imaging," *International Journal of Information Technology Convergence and Services,* vol. 4, 2014.

[15] E. E. M. Azhari, *et al.*, "Tumor detection in medical imaging: A survey," *International Journal of Advanced Information Technology,* vol. 4, 2014.

[16] N. A. Mohamad, *et al.*, "Bacteria identification from microscopic morphology using naïve bayes," *International Journal of Computer Science, Engineering and Information Technology,* vol. 4, 2014.

[17] N. A. Mohamad, *et al.*, "Bacteria identification from microscopic morphology: A survey," *International Journal on Soft Computing, Artificial Intelligence and Applications,* vol. 3, 2014.

[18] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *AAAI*, 1992, pp. 129-134.

[19] I. Kononenko, *et al.*, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Applied Intelligence,* vol. 7, pp. 39-55, 1997.

[20] N. Landwehr, *et al.*, "Logistic model trees," *Machine Learning,* vol. 59, pp. 161-205, 2005.

[21] M. Sumner, *et al.*, "Speeding up logistic model tree induction," in *Knowledge Discovery in Databases: PKDD 2005*, Ed: Springer, 2005, pp. 675-683.

[22] C. Perlich, *et al.*, "Tree induction vs. logistic regression: A learning-curve analysis," *The Journal of Machine Learning Research,* vol. 4, pp. 211-255, 2003.

[23] S. R. Hingorani, *et al.*, "Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse," *Cancer cell,* vol. 4, pp. 437-450, 2003.