Cancer Biomarkers Classification from SELDI-TOF Mass Spectrometry for Clinical Proteomics: An Approach of Dimensionality Reduction

Azween Abdullah1* and Ramachandran Ponnan2

1School of Computing and IT, Taylors University, Subang Jaya, Selangor, Malaysia. 2School of Communications, Taylors University, Subang Jaya, Selangor, Malaysia.

Authors’ contributions

This work was carried out in collaboration between both authors. Author AA designed the study, performed the statistical analysis, wrote the protocol and wrote the first draft of the manuscript. Author RP managed the analyses of the study and managed the literature searches. Both authors read and approved the final manuscript.

ABSTRACT

Cancer diagnosis from proteomic profiles has reformed the medical procedures in a significant manner with its enhanced accuracy rate as compared to other ultrasound imaging based process. Efficient classification of suitable cancer biomarkers from proteomic data helps in early diagnosis of cancerous diseases. Mass spectrometry (MS) with protein chip based technology such as the Surface Enhanced Laser Desorption and Ionization Time of Flight (SELDI-TOF) can be used for presence as well as absence of diseases by extracting protein spectra based on m/z ratio and intensity of the protein. For mass spectrometry, efficient and robust feature selection technique is required which can reduce the number of features as much as possible in less time and eliminates any irrelevant or redundant features which can affect the classification performance. This work incorporates an energy based dimensionality reduction for huge data to perform clustering ensemble binary classification of cancer and normal patterns by evaluating biomarker signatures at a higher rate of accuracy. The proposed approach has overcome existing issues especially feature

*Corresponding author: E-mail: Azween.Abdullah@taylors.edu.my;
The experimental results show that both SELDI-TOF data sets extracted from WCX2 and H4 chip can be classified with only two features at relatively higher rate with negligible false alarms.

Keywords: Energy-based dimensionality reduction; biomarker signatures; clustering ensemble binary classification.

1. INTRODUCTION

Cancer causalities are increasing rapidly worldwide and various medical data are under consideration to take possible measures for early stage cancer biomarker identification [1,2]. Apart from ultrasound imaging for the organs, some high dimensional data from sources such as Mass Spectrometry (MS) or microarray can also be utilized which can be effective, if suitable techniques and methods are incorporated [3,4]. Earlier studies have shown that molecules concentration within the blood can help technologist to determine their mass spectrum data for cancer type and relative complexities [5]. Biological technologies for mass spectrometry namely Matrix assisted laser desorption and Ionization -Time of flight (MALDI-TOF) identifies possible positioning from proteomic data required to diagnose various cancer disease which requires much time and procedures in other clinical examinations based on imaging and laboratory test samples [6,7]. In all other aspects, MALDI-TOF and SELDI-TOF have very similar processing times. Alternatively, SELDI-TOF based data can also be extracted by utilizing Weak Cation Exchange (WCX2) or H4 protein chip, in low as well as high resolution, for cancer diagnosis.

The MS data can be classified with highly developed diagnostic systems but requires large number of features, which needs to be overcome by analyzing the discrimination power of the features for possible classes [8]. This can be done with iterative processes to locate an optimal feature subset which can reach higher detection rate. For example certain feature have clear difference in the values of cancer and normal sample, therefore such feature can be valuable in classification rather than a feature which has no significant difference between the records of known classes. Computer based diagnosis depends upon the algorithms used for analyzing and processing mass spectrometry features and associated classifier which need to adjust as much as possible to attain a higher detection rate of cancerous and normal proteomic profiles.

The dimensionality of mass spectrometry data affects the overall classification performance in terms of classifier’s overhead and time complexity[9]. Appropriate feature transformation or selection techniques should be incorporated with the classifier to enhance the efficiency measures of the system [10]. Many earlier approaches have been applied for handling features dimension by either reducing with transformation methods or with feature extraction to select only few sensitive features with higher potential to be used for distinguishing cancerous spectrum data from normal ones [11,12]. These processing can be done prior to classification with filter methods or either by integrating wrapper methods with classifier.

In recent years, there has been significant improvement in early diagnosis of cancerous biomarkers from high dimensional mass spectrometry data [3,13]. Even then many conventional methods for reducing feature dimension have suffered because of reduced detection rate or high classification overhead in processing irrelevant or redundant features if they are not properly optimized.

For discovery of biomarker element, cancer tissues with abnormal factor in the protein can be diagnosed by advanced medical procedures namely SELDI-TOF [14]. The mass spectrometry data is obtained by taking patient’s blood sample. This high dimensional spectrum data consists of above 15000 features (mass/charge) collected from respective m/z intensity value [15]. The advantage of having proteomic data fetched from mass spectrometry is its extensive details regarding abnormal signal profiling in comparison to traditional biomarkers.

The main limitation for earlier feature processing techniques involving extraction and subset selections is neglecting the interdependency of the features, their correlations which need to be resolved to get an efficient system for cancer diagnosis [16]. Most methods work solely on analyzing all the features individually to evaluate their scores or weight-age and then proceed to
use them in variable subset sizes to assess the systems performance.

To overcome such issues, this work proposes an optimized statistical metric and energy based algorithm in such manner to calculate the relative energies and dependencies of all features along with their discriminatory values by integrating the classification procedure. This technique is time consuming in the selection of various separate subsets and classification; instead the classification should start from features with higher energy level and then to combine the second best feature only when required to increase the possibility of attaining accuracy rate. Here no threshold or specific parametric adjustment is required as the only considering factor is system accuracy acquired from the optimal features.

This paper is structured as follows: Section 2 provides review of earlier approaches on feature processing and classification algorithms. Section 3 presents the material used and proposed methodology. Section 4 explains the results obtained and possible contributions. Section 5 concludes with future directions for the work.

2. EARLIER STUDIES ON PROTEOMIC DATA IN CANCER BIOMARKER DISCOVERY

Various data mining techniques have been utilized to achieve better accuracy rate by reducing the dimension of these proteomic data sets. Many feature processing models have been designed which are based on wrapper or either filter in combination with efficient classifiers to speed up the process of identification of biomarkers signatures. In this section, review of various existing feature processing and classifiers for proteomic data is provided.

In an earlier work [17], OC-WCX2 mass spectrum data set was used for experiments with classifiers like Support Vector Machines (SVM), decision trees and neural networks. Their approach also involved feature selection by statistical methods based on mean and standard deviation. They scored accuracies between 88 and 99% for different feature subsets ranging from 8 to 16 attributes with varying minimal distances. Later random projection (RP) was used in comparison with Principal component analysis (PCA) for classifiers such as Support Vector Machines, neural network and K-nearest neighbor [18]. They proved that usage of RP can enhance the effect of PCA as a feature processing for mass spectrometry data sets namely OC-WCX2a, OC-WCX2b and PC-H4 by scoring an accuracy of 95%, 99.89% and 99.90% respectively.

Plant et al. [11] proposed a feature selection method which was designed by utilizing the benefits of filter and wrapper procedures. They applied this approach for two SELDI-TOF data sets and achieved good results. They proved that modified binary search is effective on select minimal feature subset by adjusting misleading or unavailable details from the data. Support Vector Machines (SVM) classifier reached 97.8383% with 164 features and 100% for 9 features for prostate and ovarian cancer data sets respectively.

In another approach, Yu et al. [19] used SVM in a k-fold cross validations with processing methods involving binning, wavelet transform and handling coefficients variation. The sensitivity and specificity scored were 97.38 % and 93.30% for ovarian data set. Multilayer perceptron (MLP) was used to classify prostate data set by using dimension reduction techniques and reached an accuracy of 95% with only 10 features. They also used Probabilistic Neural Network (PNN) [20] and scored accuracies ranging 91% to 97% with different feature subsets extracted by noise extraction in terms of global and local vectors. A classification accuracy of 0.99% was achieved on SELDI-TOF ovarian cancer dataset by using t-test and refining for feature space reduction with kernel based classification and regression method [21]. Another approach for SELDI-TOF data set was proposed by Thakur et al. [22] involving genetic algorithm and statistical methods for selecting appropriate features. Feed Forward Neural Networks classified with an accuracy of 99.16% for cancerous and 98.50% for controlled patterns.

A neuro fuzzy classifier method was applied to reduce error rate in combination with genetic algorithm which overall reduced the feature space by selecting only 20 attributes for cancer identification [23]. Wavelet transform was later combined with kernel partial least square (KPLS) for dimension reduction and filtering to select high variant coefficients. T- score was used to calculate the corresponding means and deviations of the classes to select only relevant features by adjusting a threshold [24]. The mass spectra data set used for this work was PC-H4
with SVM and k-NN classifiers and the maximum accuracies acquired were 95.8% and 95.9% respectively.

Assawamakin et al. [25] suggested applying a two-step design where naïve bayes were used to rank the features based on their coefficients and then performed classification on potential features with hidden naïve bayes algorithm. This system was then tested with different bioinformatics (omics) data types such as SELDI-TOF, microarray and single nucleotide polymorphism microarray (SNParray). Their method worked by minimizing the number of biomarkers used and time factor. The overall accuracies achieved were 0.86 and 0.98 for prostate and ovarian SELDI-TOF data sets with only 8 features.

In 2013, Ao Kong et al. [31] worked on SELDI-TOF data sets by classifying with biomarker signature discovery method designed from SVM and k-NN (k-Nearest Neighbors) on normalized data by smoothing and noise reduction [26]. They get maximum detection rate of 95% for cancerous records and 94% for normal data records by adjusting parameters for window sizing. Seddik et al. [5] integrated wavelet and fourier transformations to separate control and cancerous records with help of neural network classifier. The system was tested with WCX2 data set reached higher accuracy with wavelet transform equivalent to 95% whereas Fourier achieved 94%. Linear discriminant analysis was also used to enhance Fourier performance but could not exceed 80.7% accuracy rate.

Principal component analysis (PCA) was combined with linear discriminant analysis (LDA) to perform classification on principal components with high variance. This ensemble approach increased the performance rate up to 7% and 18% in sensitivity and specificity respectively than LDA on ovarian data set [27]. Same approach of PCA and LDA was used earlier as well by Lilien et al. [28] with a probabilistic algorithm Q5 and achieved an accuracy of 97% for the SELDI-TOF data set.

In another work [29], derivative component analysis (DCA) was used to handle proteomic data in an implicit multi resolution method of derivatives by removing noise. For classification SVM was applied to perform linear separation by blemishing reproducibility problems. Matrix assisted laser desorption-Time of flight (MALDI-TOF), Surface enhanced laser desorption and ionization – time of flight (SELDI-TOF) and quadrupole time of flight (SELDI-QqTOF) data profiles were utilized for the experiments. The results achieved were better than traditional SVM as DCA extraction was used to de-noise the data. They also explained that using SVM kernels especially Radial basis function (RBF) causes over fitting on such high dimensional data. The system reached 99.52% accuracy with 100% sensitivity and 99.17% specificity on ovarian data set.

This research differs from the earlier works in that it includes an energy based dimensionality reduction for huge data to perform clustering ensemble binary classification of cancer and normal patterns by evaluating biomarker signatures at a higher rate of accuracy. All of the methods above use parametric evaluation with a single feature selection and classification techniques.

3. MATERIALS AND METHODS

This section covers the methodology proposed for the classification of SELDI-TOF data sets. Fig. 1 represents the phases involved in the system design. The classification approach used is k-fold cross validation instead of conventional training and testing process which allows optimizing the system’s parameters as much possible to attain higher accuracy.

![Fig. 1. Design of diagnostic system for cancer biomarkers](image)

3.1 SELDI-TOF Datasets

There are two different mass spectrometry datasets from Clinical Proteomics Programs Databank, National Institute of Cancer [30] used...
for proposed classification method of controlled and cancer samples. First dataset is ovarian dataset, extracted by utilizing Weak Caution Exchange (WCX2) protein chip along with high quality PBSII SELDI-TOF mass spectrometer for producing the mass spectra [18]. There are around 91 controls and 162 cancer samples which are not shuffled yet and will be processed using feature metrics and then classification. The second dataset is prostate dataset which is produced differently by adapting Hydrophobic (H4) protein chip and PBS1 SELDI-TOF mass spectrometer [31]. The chip was designed as per preferred methods. This was done to attain sufficient low molecular weight details to be beneficial for identifying cancerous samples. There are overall 63 control and 69 prostate cancer samples. Both datasets have 15154 features representing values at particular cancer samples. Both datasets have 15154 features representing values at particular molecular mass-to-charge (m/z) intensities [5] along with a label indicating feature denoting as control or cancer with number 1 and -1 respectively.

3.2 Multivariate Distribution for Dimensionality Reduction

For the purpose of reducing the number of features, energy based feature selection mechanism is adopted to choose the most significant features for classification, i.e. the mixture of features that will best distinguish the cancerous from the control (normal) classes. To accomplish this, a two-step feature selection process is performed.

3.2.1 Statistical energy

The statistical energy is calculated as follows:

\[
\text{Energy}_{\text{metric}}(k) = \frac{\sum_{i}^{n} \sum_{j}^{n_i} |n_i^j(k)|}{\sum_{i}^{n} n_i(k)},
\]

where \(n_i^j(k)\) is feature \(k\) of sample \(j\) in class \(i\), \(n\) is the number of classes; \(n_i\) is the number of images in class "\(j\)".

The null hypothesis \(H_0\): Energy of feature "\(k" = 0\), is indeed need to be tested against the alternative hypothesis, \(H_1\): Energy of feature "\(k\) ≠ 0."

After calculating the energy, certain threshold is applied and all values that are below that threshold are removed.

3.2.2 Statistical metric

This can be calculated as follows:

Assume \(m_1, m_2\) and \(m_3\) are the mean of class1, class2, and class3, respectively and \(m_T\) is the mean of all classes.

\[
\text{Let } m_T(k) = \frac{\sum_{i=1}^{n} m_i(k)}{n}
\]

Where \(cn\) = number of classes, so

\[
\text{var}_{\text{mod}}(k) = \frac{1}{n} \sum (m_i(k) - m_T(k)) \quad (2)
\]

however, \(\sum (m_i(k) - m_T(k))^2\) is not adequate to measure the classification contribution of the coefficients because it may provide similar values in the two cases. For that reason, there exists a need to present a different metric to measure the coefficients’ contribution. The metric is as follows:

\[
\text{Var}_{\text{mod}}(k) = \min_{i \neq j} \sqrt{\frac{S_i^2(m_j^k)}{n_j}} = \min_{i \neq j} |m_i(k) - m_j(k)| \quad (3)
\]

where \(S_i\) is the statistical metric of class \(i, m_j\), is the mean of class \(i\), and \(n_j\) is the number of classes. \(S_i\) can be calculated as follows:

\[
S_i^2 = \frac{\sum_{j=1}^{n_j} (x_j^i(k) - m_j(k))^2}{n_i - 1} \quad (4)
\]

where \(i = 1, 2, 3, ..., n, \) and \(n_i\) is the number of the features in class \(i\).

3.3 Proteomic Data Classification with Cluster K-Nearest Neighbor (C-K-NN)

C-K-NN is a classification algorithm that combines K-means modified algorithm and K-Near Neighbor (K-NN). This classifier was introduced by Rawat et al. [23]. The classifier is described further below.

Every single class, \(C_i\) is should become a cluster to various subclasses, \(C_{ij}\), with \(1 \leq j \leq 1\), and each subclass will be represented by its mean, \(\mu_{ij}\). Hence, the cluster analysis determines a set of groups, which will decrease the within-group variation and increase the between-group
variation. K-means clustering algorithm is applied for every class for clustering purposes, and then both the number of subclasses for every class and the initial k-vectors to initialize the K-means cluster algorithm are defined to find the ideal number of the subclasses. The number of subclasses is iterated beginning with 1 and the iteration process should stop by the following two conditions:

- All the representatives, \( \mu_{ij} \), should be close with respect to the metric \( d \) of their classes, \( C_i \) (i.e., if we classify all the representatives, \( \mu_{ij} \), we have found 100% accuracy). If there are some misclassifications of \( \mu_{ij} \), we have to decrease the parameter \( \alpha \) by multiplying it with another factor, \( \alpha' \), which is less than 1.
- The statistical metric of each class \( C_i, var_i \), does not decrease significantly in comparison to the previous iteration.

We may use \( \frac{\Delta Var}{Var} \leq \alpha \) as a criterion to quantify, if there is a decrease or it still approximately remains constant. In certain cases, it is better to stop the iteration if the condition \( \frac{\Delta Var}{Var} \leq \alpha \) has been checked twice or more (i.e., after which the statistical metric will be smoothed).

For initialization of the K-means cluster algorithm in general, we choose k-vectors, which belong to our classes’ data. This will therefore make the algorithm unstable in the sense of the final variance:

\[
var(C_i) = \sum_{j=1}^{m_i} var(C_{ij})
\]

which depends on the initial vectors.

From here, the question “How to choose the initial vectors in order to find a minimal variance?” arises. To answer this, in this paper, we have developed two algorithms: The Hierarchical Near-to-Near and Hierarchical Near-to-Mean algorithms, which might require some modification for different applications.

### 3.4 Hierarchical Near-to-Near Algorithm

This algorithm consists of calculating the distance \( d(x_{in}, x_{im}) \) for all \( x_{in} \in C_i \) and starting to cluster our classes to \( N_i - 1 \) subclasses where \( \text{card}(C_i) = N_i \). We put the two closest data in the same subclasses

\[
C_{i,1} = \{x_{in}, x_{im}\}
\]

where \( \min_{n \neq m} d(x_{in}, x_{im}) = d(x_{in}, x_{im}) \),

and we put each of the other data in separate subclasses,

\[
C_{ij} = \{x_{ij}\}, \forall j \in \{1, ..., N\} - \{n_o, m_o\}.
\]

In the next step, the following index \( n_1 \) and \( m_1 \) are considered for which

\[
\min_{n \neq m} d(x_{in}, x_{im}) = d(x_{in1}, x_{im1}).
\]

If \( x_{in1} \) and \( x_{im1} \) belong to the same subclass, \( C_{ix} \), we shall split this subclass into two other subclasses:

\[
C_{ix+1} = C_{ix} - \{x_{in1}, x_{im1}\}
\]

(6)

\[
C_{ix} = \{x_{in1}, x_{im1}\}
\]

(7)

In the case where \( x_{in1} \) and \( x_{im1} \) belong to two different subclasses, \( C_{ix1} \) and \( C_{ix2} \), respectively, we shall put \( x_{in1} \) in the subclass, \( C_{ix2} \), if \( \text{card}(C_{ix2}) > \text{card}(C_{ix1}) \) and we may put \( x_{im1} \) in the subclass, \( C_{ix1} \), if \( \text{card}(C_{ix2}) \leq \text{card}(C_{ix1}) \). Indeed, to get the cardinality of the set we may use the distance between the vectors to the set as \( d(\text{vector}, \text{mean of set}) \). When we obtain \( k \) subclasses, we stop the iteration, and our initial k-vectors will be the mean of each subclass.

### 3.5 Hierarchical Near-to-Mean Algorithm

This algorithm is almost the same as the Hierarchical Near-to-Near algorithm, except we will deal with the mean of subclass \( C_{ix} \) in the processing. We start by splitting our class \( C_i \) into two subclasses:

\[
C_{i,1} = \{x_{in}, x_{im}\}
\]

(8)

and

\[
C_{i,2} = \{x_j \mid j \notin \{n_o, m_o\}\},
\]

(9)

where \( d(x_{in}, x_{im}) = \min_{n \neq m} d(x_{in}, x_{im}) \).
We then updated our classes, $C_i$, by replacing $x_{i,n_0}$ and $x_{i,m_0}$ with their average,

$$C_i^1 = \{ \ldots, x_{i,n_0−1}, S_{i}, x_{i,n_0+1}, \ldots, x_{i,m_0−1}, S_{i}, x_{i,m_0+1}, \ldots \}$$

Where $S_i = (x_{i,n_0} + x_{i,m_0})/2$

Next, we consider $x_{i,n_1}$ and $x_{i,m_1}$ as:

$$d(x_{i,n_1}, x_{i,m_1}) = \min \{d(x_{i,n}, x_{i,m}) | d(x_{i,n}, x_{i,m}) \neq 0\}$$

We then replace $C_i^1$ and all the data in $C_i^1$ that are equal to $x_{i,n_1}$ or $x_{i,m_1}$ by $S_1$, which is the mean of the union of the two subclasses where $x_{i,n_1}$ and $x_{i,m_1}$ belong to:

$$S_1 = \frac{C_{n_1}x_{i,n_1} + C_{m_1}x_{i,m_1}}{C_{n_1} + C_{m_1}}$$

where $C_{n_1}$ is the number of the repetitions of $x_{i,n_1}$ inside of $C_i^1$, and $C_{m_1}$ is the number of the repetitions of $x_{i,m_1}$ inside of $C_i^1$. Our algorithm stops once the number of distinct vectors inside of $C_i^1$ is equal to $k$. Our classification algorithm does not need to keep all the data, only the average of each subclass. This is the outstanding feature of this new clustering. To classify a new data or vector $x$, we use k-NN algorithm, i.e., we assign $x$ to the class $C_i$ for which:

$$i = \arg \min_{j} d(x, \mu_{i,j})$$

where $\arg \min_{j} d(x, \mu_{i,j}) = i_0$

4. EXPERIMENTAL RESULTS

For this experiment study, two mass spectrometry data sets generated by SELDI-TOF are used to testify the proposed system. The main contribution of this work is to select a few optimum features which would have the high discriminatory power to classify data in known classes i.e. cancer and normal. One data set is ovarian cancer data (OC-WCX2b) and the second data set is prostate cancer data (PC-H4). Both are high dimensional data set as they include 15154 features along with a label feature indicating the pattern as cancerous or controlled. Next, the details of the data sets are provided in the Table 1.

The first step for both data sets is to calculate the energy metrics for the features and arrange them in ascending order to further proceed for elimination of irrelevant features. Figs. 2 and 3 shows the energy distribution of the features from the selected data sets. The energy level is quite low for the first 40% of the features (m/z values) and they would not be beneficial to use for classification of biomarkers as cancerous or controlled. Such features will be eliminated in the feature processing phase, so that the further steps will be executed in identifying the optimal feature subsets for cancer diagnosis.

![Fig. 2. Energy metric of OC-WCX2 data set](image1)

![Fig. 3. Energy metric of PC-H4 data set](image2)

| Data set | No of features | No of Patterns (Normal/ Abnormal) | Train/Test Ratio (%) |
|----------|----------------|-----------------------------------|----------------------|
| Ovarian  | 15154          | 91/162                            | 70/30                |
| Prostate | 15154          | 63/69                             | 70/30                |
The energy distribution is relatively same for both data sets, as they have the same number of features based on mass/charge (m/z) values ranging from \(-7.86 \times 10^{-5}\) to 19995.51 and their corresponding intensity value for the patients using the WCX2 and H4 chips.

Then for evaluation of the best features which have maximum effect on the classification, Multivariate T distribution is used. This procedure is done in incorporation with the classifier to determine the features which will have maximum effect on accelerating accuracy. In the beginning, all the patterns are classified by utilizing one best feature as this process takes less time and is more convenient to determine the most appropriate feature for the biomarker identification. The features with higher energy metric are trained and tested by the system and their corresponding performance measures such as Accuracy (red color) and False positive (green color) and False negative (blue color) is calculated. The resultant performance for individual feature based biomarkers classification is presented in the Figs. 4 and 5. This technique helps in identifying the potential features for the classification.

In the next phase of classification, the feature subsets used in the experiments are based on two features with high energy coefficients to enhance the detection accuracy and reduce the false alarms. Figs. 6 and 7 shows the corresponding performance measures such as Accuracy (red color) and False positive (green color) and False negative (blue color) for both data sets.
The acquired accuracy is 100 with no false alarms by using the feature set of two features for both data sets.

Here the significant point is both mass spectrometry data sets have the same accuracy in this phase with quite similar results as seen in the figures which shows the consistency of features and the selected technique.

The major difference in both data sets is the feature ranking order which is dissimilar even though the same feature type was used in the experiment. This shows that both cancers need different feature (mass/charge) value to diagnose the particular cancer type. So instead of using a feature dimension of around 15000, it’s prudent to use only the cancer relevant features depending upon the cancer type to get the desired results.

The main contribution of this work is to identify specific potential mass/charge which can be used for cancer diagnosis (ovarian or prostate) from collected proteomic data from mass spectrometry technique.

The top ten features (mass/charge) for OC-WCX2 data set are presented in Fig. 8. Feature number 1925 can score an accuracy of 94.6% individually and can enhance the accuracy up to 100% by forming an optimal feature subset with another feature number 1924.

The first 5 features for OC-WCX2 data set are in the same range of 1921 to 1925 with a slight difference of their energy level. Another important finding is that neither a high mass/charge nor lower mass/charge is effective for ovarian cancer identification and classification from the controlled cases.

Whereas, the top ten features (mass/charge) for PC-H4 data set are completely different as presented in Fig. 9. Feature number 1680 can score an accuracy of 99.2% individually and can enhance the accuracy up to 100% by forming an optimal feature subset with another feature number 1678.

As discussed in Section 2, many earlier techniques have been applied to reduce minimum features for cancer detection apart from maximizing the accuracy rate. The comparison of proposed feature reduction algorithm with existing approaches is presented in Table 2.

This work has significantly reduced the features space compared to earlier works that have required at least 8 features. The proposed statistical method calculates the energy metric and then identifies the minimum possible feature subsets by the classification process. Such optimized diagnostic systems are essential for timely detection of real-time clinical procedures for effective and timely medical interventions.
Fig. 9. Optimal features ranking for PC-H4 data set

Table 2. Performance comparison with earlier works

| Author (Year) | OC-WCX2 data set                  | PC-H4 data set                  |
|---------------|-----------------------------------|---------------------------------|
| Hong, et al. [17] | 88-99% accuracy rate for 8-16 features | -                              |
| Plant, et al. [11] | 100% accuracy for 9 features     | 97.83% accuracy for 164 features |
| Bougioukos, et al. [20] | 98% accuracy for 8 features      | 95% accuracy for 10 attributes  |
| Assawamakin, et al. [25] | 94.5% accuracy for 1 features | 86% accuracy for 8 features     |
| Proposed Work | 100% accuracy for 2 features      | 100% accuracy for 2 features    |

5. CONCLUSION

Analysis of proteomic profiles for clinical diagnosis of cancer biomarkers demands efficiency and procedural optimization. Mass spectrometry based proteomic data can be useful in the identification of cancer stages by analyzing the possible discrimination among them with ensemble machine learning algorithms. In this work, statistical metric and energy method has been incorporated with Clustered k-NN classifier to reduce the dimension of the high resolution data and classify the extracted high energized features. The results showed significant reduction in the dimension by utilizing even a single particular m/z intensity feature to distinguish both prostate and ovarian cancer from the normal data. This approach has also decreased the time factor by selecting features based on their classification accuracy and energy level which optimized the overall biomarker identification process. This technique can be further applied for many other cancerous biomarkers to eliminate the dimensional issues in proteomic and genomics datasets.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Rose RJ, Allwin S. Computerized cancer detection and classification using ultrasound images: A survey. International Journal of Engineering Research and Development. 2013;5(7):36-47.
2. Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: Opportunities and potential limitations. Mol Cell Proteomic. 2004;3(4):367-78.
3. Mesri M. Advances in proteomic technologies and its contribution to the field of cancer. Advances in Medicine; 2014;25. Article ID 238045. Available: http://dx.doi.org/10.1155/2014/238045
4. Nie F, Huang H, Cai X, Ding CH. Efficient and robust feature selection via joint $\ell_2,1$-norms minimization. Advances in Neural Information Processing Systems. 2010; 1813-1821. Available: https://papers.nips.cc/book/advances-in-neural-information-processing-systems-23-2010
5. Seddik AF, Hassan RA, Fakhrdein MA. Spectral domain features for ovarian cancer data analysis. The 2006 International Conference on, 2006;313-
319. Journal of Computer Science (JCS). 2013;9.
6. Pyatnitskiy M, Karpova M, Moshkovskii S, Lisitsa A, Archakov A. Clustering mass spectral peaks increases recognition accuracy and stability of SVM-based feature selection. Journal of Proteomics & Bioinformatics. 2010;3(2):48-54.
7. Gemoll T, Roblick UJ, Habermann JK. MALDI mass spectrometry imaging in oncology (Review). Mol Med Rep. 2011;4(6):1045-51.
8. Boja ES, Rodriguez H. Proteogenomic convergence for understanding cancer pathways and networks. Clin Proteomics. 2014;11(1):1559-0275.
9. Luke BT, Collins JR, Habermann JK, Prieto DA, Veenstra TD, Ried T. Comparing biomarkers and proteomic fingerprints for classification studies. Journal of Biomedical Science and Engineering. 2013;6(4):453-465.
10. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learning, 2002;46(1):389-422.
11. Plant C, Osl M, Tilg B, Baumgartner C. Feature selection on high throughput SELDI-TOF mass-spectrometry data for identifying biomarker candidates in ovarian and prostate cancer. Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops @ Institute of Biomedical Engineering University for Health Sciences, Biomedical Informatics and Technology (UMIT), Hall in Tirol, Austria; 2006. Available: https://pdfs.semanticscholar.org/0a39/7732fdacb9092c935bd1536c483036bd53.pdf
12. Krier C, Francois D, Wertz V, Verleysen M. Feature scoring by mutual information for classification of mass spectra. Applied Artificial Intelligence: Proceedings of the 7th International FLINS Conference, Genova, Italy. 2006;557-564. Available: https://perso.uclouvain.be/michel.verleysen/papers/flins06ck.pdf
13. Levner I. Feature selection and nearest centroid classification for protein mass spectrometry. BMC Bioinformatics. 2005;6-68. Available: http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-68
14. Pham T. Mass spectrometry based cancer classification using fuzzy fractal dimensions. progress in pattern recognition, image analysis and applications. Rueda L, et al. Eds., ed: Springer Berlin Heidelberg. 2007;4756: 614-623. Available: http://link.springer.com/chapter/10.1007/978-3-540-76725-1_64#page-1
15. Tessitore A, Gaggiano A, Cicciarelli G, D. Verzella, Capece D, Fischietti M, Zazzeroni F, Alesse E. Serum biomarkers identification by mass spectrometry in high-mortality tumors. International Journal of Proteomics; 2013. Available: https://pdfs.semanticscholar.org/f007/b93b422e2c2196c86be1b1b247a5b963ffac.pdf
16. Yang P, Zhang Z, Zhou B, Zomaya A. A clustering based hybrid system for biomarker selection and sample classification of mass spectrometry data. Neurocomputing. 2010;73(13-15):2317-2331.
17. Tang H, Mukomel Y, Fink E. Diagnosis of ovarian cancer based on mass spectra of blood samples. IEEE International Conference. Systems, Man and Cybernetics. 2004;3444-3450. Available: https://www.cs.cmu.edu/~eugene/research/full/mass-spectra.pdf (Cited 9 January 2017)
18. Loy C, Lai W, Lim C. Dimensionality reduction of protein mass spectrometry data using random projection. Neural Information Processing. 2006;776-785. (Cited 9 January 2017)
19. Yu J, Ongarello S, Fiedler R, Chen X, Toffolo G, Cobelli C, et al. Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. Bioinformatics. 2005; 21(10):2200-2209.
20. Bougioukos P, Cavouras D, Daskalakis K, Kostopoulos S, Kaltzis I, Nikiforidis G, et al. Proteomic mass spectra classification for biomarker discovery in prostate cancer employing pattern recognition techniques. 2nd International Conference on Experiments/Process/System Modelling/ Simulation & Optimization. Athens: IC-EpsMsO; 2017. Available: http://www.bme.teiath.gr/medisp/pdfs/BOUGIUKOS_2007_EpsMsO_Proteomic%20Mass%20Spectra.pdf (Cited 9 January 2017)
21. Tang K, Li T, Xiong W, Chen K. Ovarian cancer classification based on
dimensionality reduction for SELDI-TOF data. BMC Bioinformatics. 2010;11(1):109.

22. Thakur A, Mishra V, Jain S. Feed forward artificial neural network: Tool for early detection of ovarian cancer. Scientia Pharmaceutica. 2011;79(3):493-505. Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3163368/

23. Rawat K, Burse K. A soft computing genetic-neuro fuzzy approach for data mining and its application to medical diagnosis. International Journal of Engineering and Advanced Technology. 2013;3(1):409-411.

24. Taşkin V, Doğan B, Ölmez T. Prostate Cancer Classification from mass spectrometry data by using wavelet analysis and kernel partial least squares algorithm. International Journal of Bioscience, Biochemistry and Bioinformatics. 2013;98(2):98-102.

25. Assawamakin A, Prueksaroot S, Kulawonganunchai S, Shaw P, Varavithya V, Ruangrjaipakorn T, et al. Biomarker selection and classification of “-Omics” data using a two-step bayes classification framework. BioMed Research International. 2013;2013:1-9.

26. Ao K, Gupta C, Ferrari M, Agostini M, Bedin C, Bouamrani A, et al. Biomarker signature discovery from mass spectrometry data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2014;11(4):766-772.

27. Begam BKumar J. Classification of mass spectral data using linear discriminant analysis with dimensionality reduction. International Journal of Business Intelligents. 2014;3.

28. Lilien R, Farid H, Donald B. Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. Journal of Computational Biology. 2003;10(6):925-946.

29. Han H. Derivative component analysis for mass spectral serum proteomic profiles. BMC Medical Genomics. 2014;7(Suppl 1):S5.

30. Institute N. Clinical proteomics program databank - proteomic patterns. Home.Ccr. Cancer.Gov.; 2017. Available:https://home.ccr.cancer.gov/ncifdproteomics/opatterns.asp

31. Kong A, Gupta C, Ferrari M, Agostini M, Bedin C, Bouamrani A, et al. Biomarker Signature Discovery from mass spectrometry data. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2014;11:766-772.

© 2016 Abdullah and Ponnan; This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/17575