GRC-MS: A GENETIC RULE-BASED CLASSIFIER MODEL FOR ANALYSIS OF MASS SPECTRA DATA

Sara Al-Osimi[1] and Ghada Badr[2]

[1]Department of Computer Science, Shaqra University, Riyadh, KSA
[2]IRI - The City of Scientific Research and Technological Applications, Alex, Egypt

ABSTRACT

Many studies uses different data mining techniques to analyze mass spectrometry data and extract useful knowledge about biomarkers. These Biomarkers allow the medical experts to determine whether an individual has a disease or not. Some of these studies have proposed models that have obtained high accuracy. However, the black-box nature and complexity of the proposed models have posed significant issues. Thus, to address this problem and build an accurate model, we use a genetic algorithm for feature selection along with a rule-based classifier, namely Genetic Rule-Based Classifier algorithm for Mass Spectra data (GRC-MS). According to the literature, rule-based classifiers provide understandable rules, but not accurate. In addition, genetic algorithms have achieved excellent results when used with different classifiers for feature selection. Experiments are conducted on real dataset and the proposed classifier GRC-MS achieves 99.7% accuracy. In addition, the generated rules are more understandable than those of other classifier models.

KEYWORDS

Mass spectrometry, data mining, biomarkers, rule-based classifier, genetic algorithm.

1. INTRODUCTION

Mass spectrometry (MS) is an efficient technique that has been widely used in many disciplines, such as science, engineering, and biology. Recently, MS has been used in the bioinformatics field to identify the amounts of chemical and biological materials in human tissue or serum to use later as biomarkers. These biomarkers can be used as measures for clinical assessments to monitor and predict individuals' health conditions in order to plan suitable therapeutic interventions [1]. However, because the data generated using the MS technique is so huge and extensive, it is difficult to extract any useful knowledge or biomarkers; Many studies have been done to develop data mining analysis tools (i.e., classification, clustering, correlation analysis, etc.) for the interpretation and extraction of accurate knowledge from MS data. However, the results of most of these studies have not been satisfactory. Even when the studies do achieve good results, experts may struggle to understand them. According to the literature [2], rule-based classifiers yield acceptable results when they are applied to the analysis of discrete data. In addition, these

David C. Wyld et al. (Eds) : AIAPP, CRIS, CoSIT, SIGL, SEC - 2018
pp. 15–36, 2018. © CS & IT-CSCP 2018 DOI : 10.5121/csit.2018.80402

classifiers have the unique ability to provide very meaningful outcomes. However, unfortunately, rules-based classifiers do not achieve the quality required for analysis of MS data.

In this paper, we propose an efficient and meaningful approach that uses Genetic Algorithms (GAs), namely GRC-MS, to select features and then build a rule-based classification model with the objective of classifying and understanding MS data. We also test our proposed approach on a real dataset of ovarian cancer patients in order to measure the accuracy of the proposed approach. The proposed approach is intended to be a general framework that can be used for the analysis of any MS data or related continuous data. To the best of our knowledge, the combination of rule-based classifiers with GAs as the feature selection technique has not yet been applied to MS data.

This paper is organized as follows: Section 2 provides a background about the MS techniques: preprocessing, some feature selection, and classifiers that are used for MS data. Section 3refers to some of the studies that use GA technique as a feature selection approach for MS data. In addition, it summarizes some of the studies that use rule base techniques as classifiers for MS data. Section 4 explains the steps of our proposed approach, GRC-MS. The experimental setup and results on a real dataset are presented in Section 5.Section 6, discuss the results. Finally, Section 7 concludes the paper and discusses future work.

2. BACKGROUND

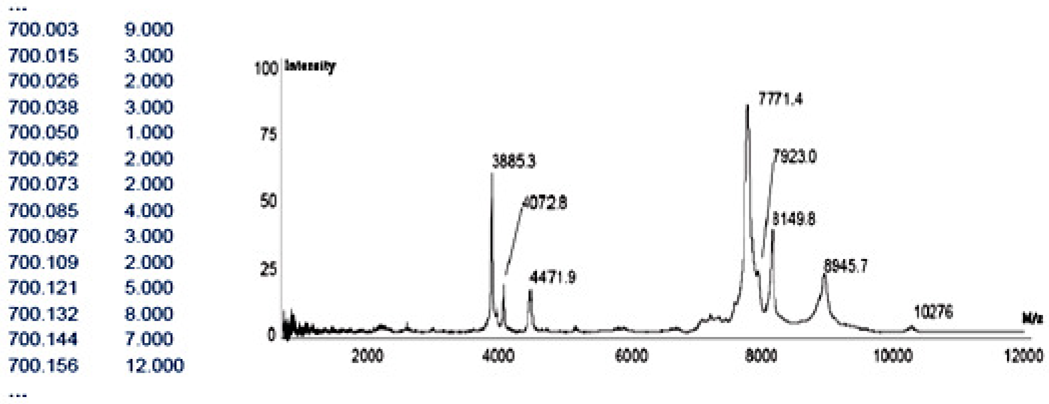
Mass spectrometry (MS) is a recently developed technique that is used to identify, analyze, and determine the elements, molecules, and atomic structures of any given sample [3]. MS quickly and accurately determines the relative numbers of molecules present in complex biological or chemical samples by transforming these samples into ions, detecting their mass-to-charge ratios (m/z), and then measuring the intensity of each ion type [4]. This technique is used primarily to study the effects of ionizing energy on sample molecules [3]. It has several beneficial characteristics, such as speed and sensitivity. Moreover, because MS has a variety of possible applications, it is preferable to other analytical methods and, as a result, has progressed rapidly over the last decade. Today, MS is used in a number of applications, such as biochemical problems, pollution control, atomic physics, food control, forensic science, reaction kinetics, geochronology, inorganic chemical analysis, process monitoring, and so on [4].

2.1 Proteomics

Proteomics, a term that is first coined by Australian scientist Marc Wilkins in 1994, is an emerging area in bioinformatics [7]. It provides information about proteins and their interactions in the human body. The major aim of most proteomic studies is the detection of proteins of interest, which are known as biomarkers. The term “biomarkers” refers to protein molecules that facilitate the detection of a particular cell type and that identify cell characteristics, such as cells’ ability to perform their functions. [8]. The discovery of biomarkers in MS data is useful for the early diagnosis of diseases. Most researchers hope to discover novel and powerful diagnostic proteomic tools to detect these biomarkers [8]. Recently, several techniques have been developed for analyzing bodily fluids, such as human serum, human urine, and, in some studies, tumor tissue, to achieve protein profiling. Commonly, the analysis of body fluids is accomplished using MS techniques [9]. Two major techniques are intended for proteomic analysis: MALDI and SELDI. MALDI-TOF MS is a new and widely used technique for discovering biomolecules, such as proteins with molecular masses between 400 and 350000 Da, in samples [6].

2.2 Mass Spectrometry

MS experiments are generally conducted in three main stages: the data generation stage, the data preprocessing stage, and the data analysis stage [4] [5]. In the first stage, MS techniques generate data that are represented as a huge sequences of pairs, called matrix, spectrum, or MS data [4]. This spectrum contains mass-to-charge ratio values and intensity values [6]. The mass-to-charge ratio values (which are represented on the x-axis) depend on the molecular mass detected in the sample, and the intensity values (which are represented on the y-axis) depend on the quantity of molecules detected in the sample (Figure 1) [6]. Depending on the resolution of the MS technique, a spectrum can contain hundreds or thousands of pair values [7]. Data preprocessing involves cleaning the data and improving their quality. On the other hand, during data analysis, data mining or pattern extraction techniques are applied to extract knowledge.



...	
700.003	9.000
700.015	3.000
700.026	2.000
700.038	3.000
700.050	1.000
700.062	2.000
700.073	2.000
700.085	4.000
700.097	3.000
700.109	2.000
700.121	5.000
700.132	8.000
700.144	7.000
700.156	12.000
...	

Figure 1. Output signals from a mass spectrometer consisting of m/z and intensity values [9].

2.3 Data Mining in Mass Spectrometry

Data mining is a well-known approach that is used in science and business to extract useful information from large and complex datasets [7][2]. The steps involved in data mining include (Figure 2) data preparation, feature selection, model development (or pattern recognition), and model assessment. The following section focuses on the basic algorithms used in data mining for application to mass proteomic data. However, as previously mentioned, MS data are high in dimensionality, and they cannot be analyzed through the direct use of data mining techniques. Preprocessing the MS data is a crucial step in improving the data quality—and, thus, improving the quality of the classifier algorithms [6].

- **Preprocessing MS Data**

MS data or spectra are commonly influenced by errors or noise that occur during the sample preparation or the insertion into the device or by noises generated by the device itself [4]. Using the raw MS data directly for the analysis process is not effective because contaminants like noise, m/z measurement errors, and matrix size affect the results [6] [7]. In addition, because of the dimensional complexity of the spectra, efficient results cannot be obtained through the direct application of data mining algorithms or pattern extraction techniques. Therefore, cleaning the MS data is critical. To achieve clean data, different preprocessing techniques are applied to the MS data before the application of any data mining technique such as reducing noise and smoothing data, normalization, data reduction by binning, peak extraction, and peak alignment. These techniques can be used alone or in combination [10].

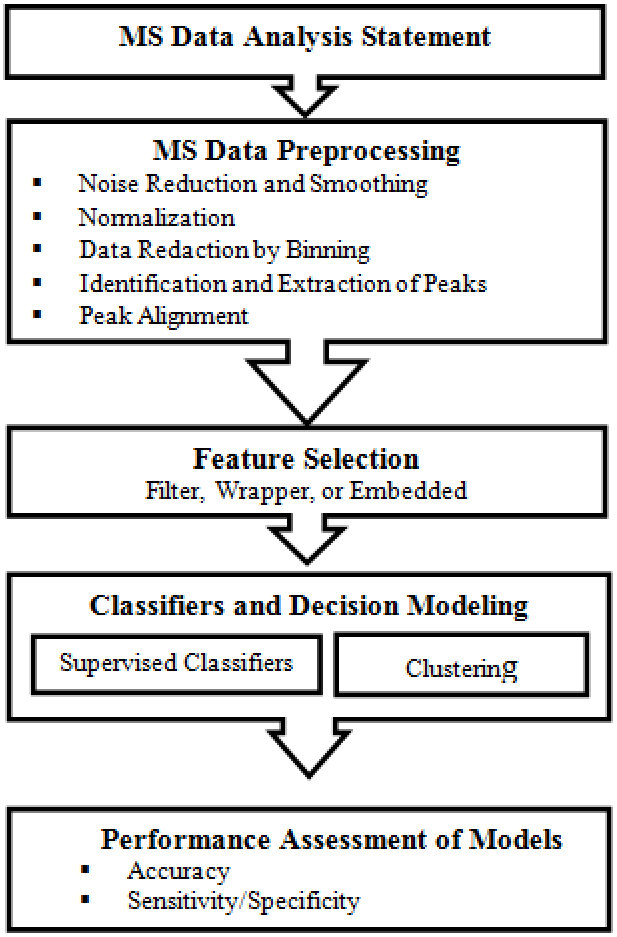

MS Data Analysis Statement
MS Data Preprocessing
Noise Reduction and Smoothing
Normalization
Data Redaction by Binning
Identification and Extraction of Peaks
Peak Alignment
Feature Selection
Filter, Wrapper, or Embedded
Classifiers and Decision Modeling
Supervised Classifiers
Clustering
Performance Assessment of Models
Accuracy
Sensitivity/Specificity

Figure 2. Typical flowchart of the critical steps in data mining and examples of the techniques available for MS data analysis.

▪ Feature Selection Techniques

The MS technique produces high-dimensional data. Compared to the number of samples, a greater number of peaks needs to be analyzed (high features-to-sample ratio datasets) [11]. Most of the problems in analyzing data stem from the size and the complexity of the datasets that are represented in tables of rows and columns. Rows represent records or cases, and columns represent data dimensions, features, and attributes [7]. In the analysis of MS data, the extraction uses the intensity of every peak in the spectrum as a feature. The number of features (or peaks) is usually large (e.g., 17,000 peaks), while the number of samples is usually small (e.g., 140 patients) [12]. However, features often contain noise with a very little or no informational value. Thus, it is necessary to select features from a large set of those likely to be useful in predicting the outputs of interest. To solve this problem, after the data is pre-processed, a feature selection phase step is performed. This step aims to detect the main parts of the spectrum that might provide a better understanding of the data’s important features, which could be used in the analysis phase [11].

Feature selection techniques can be divided into three categories [12]: filter, wrapper, and embedded. The filter technique analyzes each feature independently and eliminates features one at a time based on how they correlate with the target. Feature subsets are selected based on evaluation criterion, such as information gains. This is a simple and quick process that is sometimes referred to as independent feature selection. Moreover, filter selection methods are relatively computationally efficient [11] [12]. Examples of independent feature selection techniques used with MS or high-dimensional data include: statistical tests (i.e., t-tests [13] [14], Wilcoxon tests [17], $\chi 2$ tests [18]), information gains [19], and so on. The wrapper techniques simultaneously analyze features in groups or subsets and build the analysis model [11] [12]. Classifiers are used to assess (several) features or feature subsets. Although the process is computationally busy and potentially very time-consuming, since this technique typically requires an evaluation of every scheme at every iteration, it discovers critical information that is typically lost in independent features analysis [16] [20]. Examples of wrapper feature selection techniques that are used with MS or high-dimensional data include: genetic algorithms [21] [22], sequential searches [23], and estimations of distribution algorithms [24]. In embedded techniques, the search for a best set of features is made into the classifier construction. They learn which set of features can best contribute to the accuracy during the creation of the model [12]. These techniques make no distinction between learning and feature selection. Embedded techniques have the advantage of including the interactions with the classification model, while simultaneously being far less computationally intensive than wrapper methods [12]. Examples of embedded feature selection techniques that can be used with MS or high-dimensional data include random forest techniques [25] and weight vector support vector machine techniques [26].

- **Classifiers and Decision Models for MS Data**

For MS Data, usually classification or supervised learning uses to predict or classify new cases. Where, previous knowledge about classes can be used to classify new cases. The previous knowledge is built using a training dataset, which includes input values and their output classes. In the training stage, the training dataset is used to define how the features are to be selected and combined to distinguish among the different classes. In the testing stage, the weighted features are applied to classify a new test dataset. The test dataset’s class is not known, and the dataset has never before been seen by the model. If the model classifies new cases correctly, it is a good model. A wide range of algorithms, such as decision tree algorithms, SVMs, ANNs, and so on, have been developed for classification. In this subsection, we well indicate to some of well-known classifiers that used for MS data.

i. iDecision Tree (DT) Classifier

Decision tree (DT) a hierarchical tree structure model that is described as a set of rules, presented in a visual form that is very easy to understand. The DT was used by Vlahou et al. [28] to analyze the MS data and it is achieved 80% accuracy in discriminating between ovarian cancer patients and healthy controls. In addition, Su et al. [29] used DT to analyze the MS data and they obtained 85.3% accurate.

ii. Artificial Neural Networks (ANNs) Classifier

Artificial neural networks (ANNs) are another popular classifier used in MS data analysis. Ward et al. [13] used an ANN algorithm to analyses non-cancer and colorectal cancer samples via SELDI to identify colorectal cancer biomarkers. The ANNs with the seven highest peaks obtained 95% sensitivity and 91% specificity. Also, Chen et al. [30] used ANNs to diagnosis colorectal cancer. The proposed approach obtained 91% sensitivity and 93% specificity.

iii. Naive Bayesian (NB) Classifier

The naive Bayesian is "a simple probabilistic classifier based on the Bayesian theorem with the (naive) independence assumption" [31]. Zheng [31] compared the performance of the naïve Bayesian (NB) and the logistic regression (LR) on MS data. They found the average performance of the NB (around 90%) and the logistic regression depended on the amount of training data.

iv. Support Vector Machine (SVM) Classifier

Support vector machines (SVMs) attempt to find the best hyperplane line that separates all class A data points from class B data points. Wagner et al. [32] found that the linear SVM was the only classification method that obtained robust performance (98% accuracy). Also, Prados et al. [19] achieved 97% sensitivity and 71% specificity when used SVM-based model to classify MS data.

▪ Performance Assessment of Models

The last stage of the data mining modeling process is the assessment or validation of the model. Below, we will discuss the measures of accuracy, sensitivity, and specificity.

The classification of accuracy is calculated by "comparing the ratio of the number of correctly classified samples to the total number of samples in the test data" [33]. However, when the spread of a certain class is greater than that of other classes, the majority class will create unequal results. In this scenario, the accuracy measure will not be true. Most MS data analysis studies have used accuracy to report their results [33] [34].

There are four possible results when test decisions are built for data with two class samples: true-positive, true-negative, false-positive, and false-negative [33]. The true-positive rate is known as sensitivity. It represents the ratio of the number of correctly classified positive samples to the total number of positive samples. When the effect of incorrectly predicting a diseased person as healthy is high, high sensitivity is preferred in medical diagnoses. Specificity refers to the false-positive rate, or the probability that a healthy subject will be incorrectly classified as unhealthy [14]. When a false alarm would result in unwanted tests or treatments, high specificity is desirable here [7]. In very good classification, both sensitivity and specificity should be high, though different levels of these measures are accepted depending on the application. However, it is very hard to compare the results of different studies using only measures of sensitivity and specificity [33].Up to our knowledge, many approaches failed to achieve high accuracy. Even when high accuracy is obtained, the "black box" nature of these proposed approaches is a major issue. To address this problem and to build an accurate and understandable model, we propose to use a rule-based classifier approach along with using GAs for feature selection.

3. LITREATURE REVIEW

In this section, we explore some of the studies that use GA technique for feature selection on MS data. In addition, we illustrate some studies that use rule-based techniques as a classifier on spectrum data with classifiers.

A. Genetic Algorithm Based Feature Selection for MS Data

One popular algorithm that isused for feature selection purpose is a genetic algorithm (GA). A GA searches for optimal MS data features or biomarkers to use in the mining stage in order to distinguish patients from controls in an accurate way. Here wediscuss GA as a feature selection approach for MS data. Many studies have used GA for feature selection before applying a classifier. In 2009, Reynès et al. [35] developed a new model using a GA for feature selection and a very simple tree as a classifier. The GA in this model sought to choose a set of interesting features in a spectrum to achieve the best split points in the tree. First, the authors applied preprocessing steps to the dataset. The dataset contained 162 ovarian cancer samples and 91 control samples. Of these, 46 control samples and 81 cancer samples were randomly chosen for use as a training set; the rest (45 control and 81 cancer samples) were later used for testing. The authors obtained 98% accuracy after building the tree with three different peaks (245 Da, 434 Da, and 649 Da). The major issue in this technique when GAs return large numbers of features the DT become large and difficult to understand.

In 2004, Mohamad et al. [36] proposed a new model for applying a GA to seek and identify potential informative features using an SVM classifier. Experimental results on a breast cancer dataset (which contained 200 samples for training and 77 samples for testing) and a leukemia cancer dataset (which contained 38 samples for training and 34 samples for testing) showed the usefulness of the proposed approach for low- and high-dimension data. The authors obtained 82% accuracy for the breast cancer dataset, with 8 features, and 100% accuracy for the leukemia cancer dataset, with 50 features. In 2004, Li et al. [37] proposed a novel model used a GA for the feature selection stage and an SVM method as a classifier. The MS dataset used included 91 control samples and 162 samples from patients with ovarian cancer. Both feature selection approaches (filter and wrapper) were explored. The results showed 98% accuracy the proposed model was applied with a filter approach.

In 2002, Petricoin et al. [38] used GA for feature selection with a cluster algorithm. The proposed algorithm was applied to a training set containing 50 ovarian cancer samples and 66 control samples. The authors obtained a sensitivity of 100%, a specificity of 95%, and a rounded accuracy of 94%.In 2007 Shah and Kusiak [39] proposed a model using GA for feature selection and DT and SVM as classifiers. They applied the proposed model to three different datasets for ovarian cancer, prostate cancer, and lung cancer. The proposed model had high classification accuracy when applied to the ovarian cancer and lung cancer dataset, such that it was able to recognize the most significant features. Table1 below summarizes some of the relevant research in this field that used genetic algorithm as feature selection for mass spectrum data.

After we review some studies that using GAs for feature selection in the analysis of MS data we found that most approaches obtained a very good accuracy results. However, there are some major challenges. For example, there is no guarantee that GAs will always simultaneously find the best solution and in the same time the minimum number of discernment features. When a GA obtains a large number of features, there will be problems using certain classifiers, such as DTs. In such cases, DTs may become very large, complex, and difficult for experts to understand. Some researchers have tried to solve this problem by adding constraints to the GA. This was the case in [39], in which the authors repeated the selection process when the number of selected features was more than 100; however, this process took a long time. Moreover, in [35], the authors added a constant to the fitness function to help it select the fewest number of features possible. However, the constant did not always work in obtaining a minimal number of features.

Table.1. Some of the research using GAs as features selection for analysis MS data

Author(s)	**Year**	**disease**	**Data**	**Feature Selection Mothed**	**Data Mining Algorithm**	**Result**
Reynès et al. [35]	2009	Ovarian Cancer	▪ 253 ovarian cancer serum samples. ▪ 162 samples from patients with ovarian cancer and 91 samples from healthy patients.	Genetic Algorithm	DT	98% Accuracy
Mohamad et al. [36]	2004	Breast Cancer	▪ 200 training samples and 77 test samples.		SVM	82% Accuracy.
		Leukemia Cancer	▪ 38 training samples and 34 test samples.			100% Accuracy.
Li et al. [37]	2004	Ovarian Cancer	▪ 253 ovarian cancer serum samples. ▪ 162 samples from patients with ovarian cancer and 91 samples from healthy patients.		SVM	98% Accuracy
Petricoin et al. [38]	2002	Ovarian Cancer	▪ 216 ovarian cancer serum samples. ▪ 100 training samples and 116 test samples.		Cluster	94% Accuracy

Shah and Kusiak [39]	2007	Ovarian Cancer	▪ 253 serum samples. ▪ 135 training samples and 118 test samples.		DT and SVM	DT: 94.07% Accuracy SVM: 97.46% Accuracy.
		Prostate Cancer	▪ 136 serum samples. ▪ 102 training samples and 34 test samples.			DT: 55.88% Accuracy SVM: 67.65% Accuracy.
		Lung Cancer	▪ 181 serum samples. 32 training samples and 149 test samples.			DT: 81.88% Accuracy SVM: 98.66% Accuracy.

After excluding all 100% accurate results due to the high chance of over-fitting, we found that the best accuracy achieved was 98.66%, which was obtained by the SVM classifier. Thus, we seek to obtain a better accuracy than this one, while simultaneously building a classifier that is easy to understand. We propose the use of a rule-based classifier, which can be understandable even when GAs return large numbers of features. This is because a rules-based classifier is easier to understandable than a DT, especially with higher numbers of features. Finally, we also seek to obtain higher classifier accuracy than that achieved by the SVM.

B. Rule-Based Classifier models for MS Data

Several machine-learning classifiers, such as DTs, SVMs, and K-nearest neighbor classifiers, have been used to successfully classify MS data. These have all achieved high predictive accuracy. However, the black-box nature of these classifiers presents major issues for developers [40] [41]. By contrast, the IF-THEN rule-based classifier can obtain satisfactory predictive accuracy, while also being easier to describe and interpret by humans than other classifiers, due to its readable IF-THEN rule structure [42]. The challenge is the extraction of a small, accurate and easy-to-interpret sets of IF-THEN rules from high-dimensional MS data. In the following, we will review various studies that have used IF-THEN rule classifiers to classify of MS data. We will then discuss these papers in order to provide a simple introduction for the development of this type of classifier.

In 2006, Ressom et al. [41] proposed a novel classifier for classifying MS data using a fuzzy IF-THEN rule-based structure. For feature selection, the authors used ant colony optimization

(ACO) with an SVM. They hoped that the combination of these two methods in the feature selection step would improve the quality of the potential biomarker identification and build an accurate fuzzy IF-THEN rules classifier. The authors collected 150 serum samples of hepatocellular carcinoma (HCC) diseases that were taken from Egypt between 2000 and 2002. Of these, 78 samples were taken from patients with HCC, and 72 samples were taken from normal individuals. After they preprocessed the samples, the authors selected 100 samples randomly as a training set, including 50 samples from the HCC patients and 50 samples from the healthy individuals. The remaining samples (28 from the HCC patients and 22 from healthy individuals) were used as a testing set for performance evaluation. The authors applied a combination of ACO and SVM to extract useful biomarkers in the feature selection stage. They found six m/z candidate biomarkers, as follows: 1863.4-1871.3, 2528.7- 2535.5, 933.6-938.2, 1737.1-1744.6, 4085.6-4097.9, and 1378.9-1381.2 Da. These six m/z candidate biomarkers were used as inputs to the IF-THEN rules classifier. The prediction accuracy of this classifier was estimated using a four-fold cross-validation method. Then, the authors used the ACO algorithm to select four rules from among the 4095 candidate rules extracted from the training dataset with the candidate biomarkers. The IF-THEN rules distinguished HCC patients from controls in the testing dataset with 91% sensitivity and 89% specificity.

Assareh and Moradi [43] proposed a model that used a t-test to select the best features and a IF-THEN rules classifier to classify the MS datasets. The dataset was for ovarian cancer, and it was made available to the public through the American National Cancer Institute (NCI) website. The ovarian cancer dataset contained 253 samples, of which 91 samples came from healthy individuals and 162 came from ovarian cancers patients. Before addressing these datasets, the authors used preprocessing to clean the datasets to enhance the classifier's performance. They binned all of the M/Z points as candidate biomarkers and applied a t-test to select the best candidate biomarkers. The t-test eliminated the biomarkers that were locally correlated, since these could correspond to the same peptide. The authors found three m/z candidate biomarkers. The proposed method achieved acceptable accuracy (86.93%) compared to two classification methods: LDA (74.24%) and KNN (68.18%).

In 2011, Wang and Palade [44] proposed a new Multi-Objective Evolutionary Algorithms-based Interpretable Fuzzy (MOEAIF) model. This model used Fuzzy C-Mean Clustering-based Enhanced Gene Selection (FCCEGS) for feature selection with fuzzy IF-THEN rules to analyze high-dimensional data, such as microarray gene expressions and MS data. The proposed model was evaluated on proteomics mass spectroscopy data from an ovarian cancer dataset containing 253 samples (91 from healthy individuals and 162 from ovarian cancer patients). Some preprocessing steps were applied to the dataset. The authors extracted eight fuzzy IF-THEN rules from the dataset (average rule length of two) using six candidate biomarkers. The candidate biomarker MZ6880.2 and the feature MZ18871.5 played important roles in most of the rules. This proposed MOEAIF model achieved 63.75% accuracy. Table 2 below summarizes some of the relevant research in this field.

In reviewing the various research papers using rule-based classifier to analyze MS data, we found that the research related to these rule-based classifiers was still very active. Various researchers had tried to improve the black-box problem of most classifiers while simultaneously achieving high predictive accuracy. Each paper proposed a model for obtaining a certain number of IF-THEN rules that would be easy for experts to understand and manipulate. However, the major challenge is improving rule accuracy by finding the best set of features. Several authors have attempted to use different feature selection methods; however, up to our knowledge, none has achieved a higher classification accuracy.

Table 2. Research using IF-THEN rules as classifiers for the analysis of MS data.

Author(s)	Year	Diseases	Data	Features Selection Method	Data Mining Algorithm	Result
Ressom et al. [41]	2006	HCC	▪ 150 serum samples of HCC diseases ▪ 78 samples from patients with HCC and 72 samples from healthy patients	ACO-SVM algorithm	IF-THEN Rule-Based	91% sensitivity and 89% specificity
Assareh and Moradi [43]	2007	Ovarian cancer	▪ 253 ovarian cancer serum samples. ▪ 162 samples from patients with ovarian cancer and 91 samples from healthy patients	T- test		86.93% Accuracy
Wang and Palade [44]	2011	Ovarian cancer	▪ 253 ovarian cancer serum samples. ▪ 162 samples from patients with ovarian cancers and 91 samples from normal.	Fuzzy C-Mean Clustering based Enhanced Gene Selection method (FCCEGS)		63.75 % Accuracy

4. Genetic-Rule-Based Classifier Model For MS Data (GRC-MS): A Proposed Approach

Given MS datasets of any diseases, the GRC-MSmodel has the following input and output:

Input: MS data obtain from controls (healthy individuals) and patients.

Output: A set of rules expressed as: I⇒C, where I refers to a set of features or biomarkers and C refers to a class label (i.e., healthy or patient).

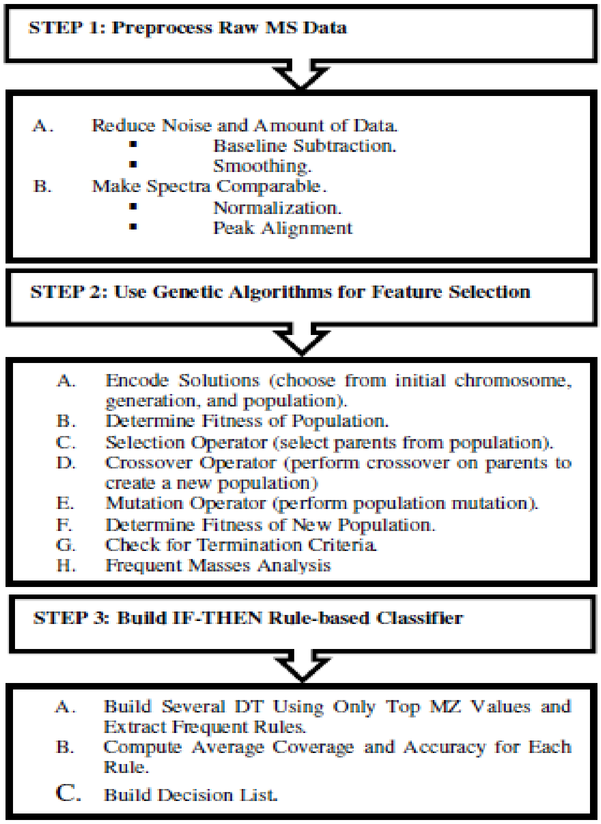

STEP 1: Preprocess Raw MS Data
A. Reduce Noise and Amount of Data.
▪ Baseline Subtraction.
▪ Smoothing.
B. Make Spectra Comparable.
▪ Normalization.
▪ Peak Alignment
STEP 2: Use Genetic Algorithms for Feature Selection
A. Encode Solutions (choose from initial chromosome, generation, and population).
B. Determine Fitness of Population.
C. Selection Operator (select parents from population).
D. Crossover Operator (perform crossover on parents to create a new population)
E. Mutation Operator (perform population mutation).
F. Determine Fitness of New Population.
G. Check for Termination Criteria.
H. Frequent Masses Analysis
STEP 3: Build IF-THEN Rule-based Classifier
A. Build Several DT Using Only Top MZ Values and Extract Frequent Rules.
B. Compute Average Coverage and Accuracy for Each Rule.
C. Build Decision List.

Figure 3. Steps of the GRC-MS model

The steps of the GRC-MSmodel are shown in Figure 3. The details of each step are explored in the following subsections:

STEP 1: Preprocess Raw MS Data

Each point on a spectrum is represented by two measurements: m/z and the intensity value. Sometimes, these points are affected or distorted by noise. Thus, preprocessing is needed to clean the MS data of noise and contaminants [9]. In addition, the preprocessing step must reduce or decrease the dimensions of the spectrum; this is important later for obtaining an efficient algorithm [33]. In this model, to correct the m/z and intensity values, we use the following steps: (A) Reduce Noise and Amount of Data and (B) Make Spectra Comparable.

A. Reduce Noise and Amount of Data

To remove a chemical noise baseline from a spectrum without harming the data is a challenging problem, since the wrong baseline correction may damage the spectrum, resulting in the wrong peak shape, peak position, or peak width [10]. We will use a function to estimate a low-frequency baseline. Then, we will subtract this baseline from the spectrum. Figures 4 show how the function corrects the baseline. These examples were taken from real dataset (ovarian cancer dataset) before and after the baseline's removal.

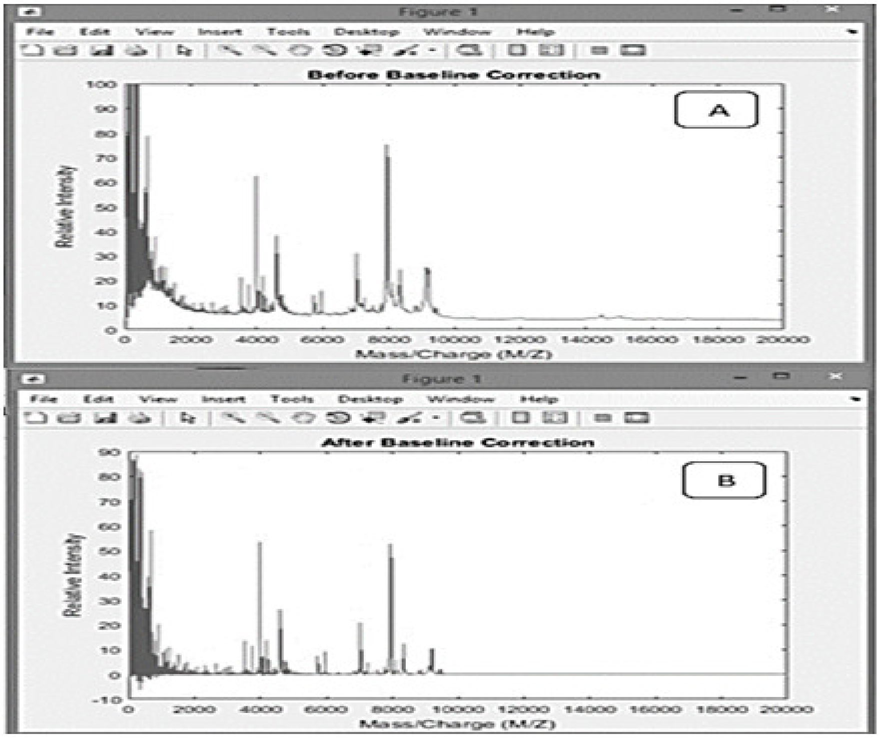


Figure 4. (A)Before baseline correction and (B) After baseline correction.

To remove electrical noise, it is important to know that spectra usually contain a combination of noises and signals. Thus, a spectrum must be de-noised to improve the validity and precision of the observed m/z values of the spectrum peaks. To accomplish this, we use Lowess smoothing and polynomial filters.

B. Make Spectra Comparable

Normalization of spectra is needed to make MS data independent of experimental differences. Normalization enables us to compare different samples, since the peak values of different spectrum fractions may be incomparable [19]. In this model, we will use the direct normalization function to calculate a re-scaled intensity value as needed. In addition, to make spectra comparable, peak alignment determines which peaks from the different spectra samples correspond to the same peak. For this, we use a sample alignment function that allows us to use a dynamic programming algorithm to assign the observed peaks in each spectrogram to the common mass/charge reference vector, if needed.

STEP 2: Use Genetic Algorithms for Features Selection

After the data preprocessing, we implement a feature selection stage, which seeks to achieve better understanding of the important features of the MS data in order to improve the classification phase later. In our model, we use GAs, which try to find optimal search solutions for problems with large datasets. Running on MS data, GAs attempt to find small sets of biomarkers that separate patient cases from control cases. This set of biomarkers, or features, is called a chromosome, such that every biomarker corresponds to a biological sample's measurements at a given m/z value (Masse). Each chromosome is evaluate by a fitness function that attempts to find the best chromosome (set of biomarkers) for separating patients from controls. GA follows the steps outlined below:

A. Encoding Solutions (choose from initial chromosome, generation, and population)

Each "chromosome" (i.e., mathematical entity, not biological) consists of d different biomarkers or features (called genes) that are initially randomly selected from all features (since most studies used all of the MS data as features after the preprocessing steps).

B. Determine Fitness of Population

The fitness value of each chromosome is determined by the chromosome's ability to classify the training set samples into patient and control groups. In our model, we will use to compute fitness values:

Fitness value = a posteriori probability + Error rate of a linear classifier.

Note: Repeat from Step C to Step G until terminated.

C. Selection Operator (select parents from population)

The chromosome with the best fitness value is entered into the next generation, and the remaining positions are filled according to the relative fitness of the chromosomes in the parent generation (probabilistically). There are many methods for selecting the best chromosomes; we are use the roulette wheel selection method, in which the parents are selected according to their fitness. Chromosomes with greater fitness will be selected more times. Thus, the better a chromosome's fitness score, the greater its chances of being selected will be.

D. Crossover Operator (perform crossover on parents to create the new population)

The crossover can be applied to either single or double points. Each gene has an equal chance of coming from either parent. Our model use single-point and fraction crossovers to determine the fraction of the next generation population created by the crossover function.

E. Mutation Operator (perform mutation of population)

When a chromosome is chosen for transmission to the next generation, a small number of genes are randomly selected for mutation (with probabilities between 0 and 1). Once the number of genes in the chromosome to be mutated has been determined, these genes are randomly selected and replaced with genes that are not already in the chromosome. In our model, we use uniform mutation.

F. Determine the Fitness of the New Population.

G. Check for Termination Criteria.

The process is terminate when a stopping criterion, such as a specific number of high chromosomes, a maximum number of generations, or a fitness value of 100%, is obtained. We use the average relative change in the best fitness function value over generations is less than or equal certain value or maximum number of generations is reached.

H. Frequent Masses Analysis

Frequency with which masses were select is then analyze. Then, using different number of masses form top frequency masses many times to dement best number of masses set which gives best rules accuracy.

STEP 3: Build an IF-THEN Rule-based Classifier

The IF-THEN rule-based classifier is built from training data using only the top selected features. Then, the IF-THEN rule-based classifier is used to predict the class label (i.e., healthy or patient) for the MS test data. The IF-THEN classification rule is as follows:

R: IF condition (C), THEN Class (C).
Example.

R1: “If biomarker 1 is less than threshold 1 and biomarker 2 is greater than threshold 2 and biomarker 3 is less than threshold 3, then the sample belongs to the patient group.”

R2 is “If biomarker 1 is greater than threshold 1 and biomarker 2 is less than threshold 2, then the sample belongs to the healthy group.”

- The LHS represent the rule condition; it is a conjunction of feature tests (biomarkers).
- The RHS denotes the rule consequent or the class label (healthy or patient).

In our work, we will build an IF-THEN rule-based classifier from a DT. In comparing the IF-THEN rule-based classifier with the decision tree, we found that the IF-THEN rule-based classifier was easier for humans to understand, especially when the DT was very huge. Then, we will assess each IF-THEN rule using rule coverage and accuracy. The Rule Ordering Scheme (i.e., Rule-Based Ordering) will then be apply. In this scheme, rule priorities are determined beforehand, and a decision list is built. This list is order according to rule quality (accuracy and coverage). The match rule that appears at the beginning of the list has the highest priority. In the event that no rule is satisfied by X, a default rule will be define for a default class, based on the training set. This class then becomes the majority class of the MS sample, encompassing all instances that are not cover by rules.

5. CASE STUDY AND RESULTS

In order to test and evaluate the accuracy of oGRC-MS model and to ensure that its rules are understandable, we apply the GRC-MS model to real data using MATLAB® software.

5.1 Dataset

We rely on open-source an MS dataset of ovarian cancer that is available to public through the clinical proteomics program of the National Cancer Institute (NCI) website(http://home.ccr. cancer.gov/ncifdaproteomics/ppatterns.asp). This dataset is labeled “Ovarian 8-7-02”. The WCX2 protein chip was used to produce this dataset. To generate the spectrum from the samples, the upgraded PBSII SELDI-TOF mass spectrometer was used. The dataset includes 162 ovarian cancer patients and 91 control (healthy) patients. The produced spectrum can be represented by a curved shape, in which the x-axis shows the m/z ratio (the ratio of the weight of a molecule to its charge) and the y-axis represents the intensity of the same molecule as a measure of the amount of that molecule. These datasets include peak intensity measurements at 15,154 points, as defined by the corresponding m/z values in the range of 0 to 20,000 Da.

5.2 Experimental Setup and Results

The following steps areapply by the GRC-MSmodel to the previous dataset:

1) Import MS data (raw data), using the xlsread or importdata function to load the data from an Excel® file. In Excel, the data are represented as discrete values, such that the rows show the m/z ratios and the columns represent the samples. The cells (the intersections of rows and columns) represented each molecule’s intensity as a measure of the amount of that molecule in the sample. After this step is finished, we have two variables loaded into MATLAB (MZ and Y). MZ is the mass/charge vector, while Y is the intensity value for all 216 patients (control and cancer).

2) Preprocess the MS data to remove all forms of noise and all artifacts introduced to the data by applying the following functions in the following order:

- msbackadj function.
- mslowess function.
- mssgolay function.
- msnorm function.

In addition, a grouping vector is created including the type of each spectrogram and the indexing vector. This “labelling” will aid in any further analysis on this dataset.

3) Run Genetic Algorithm.

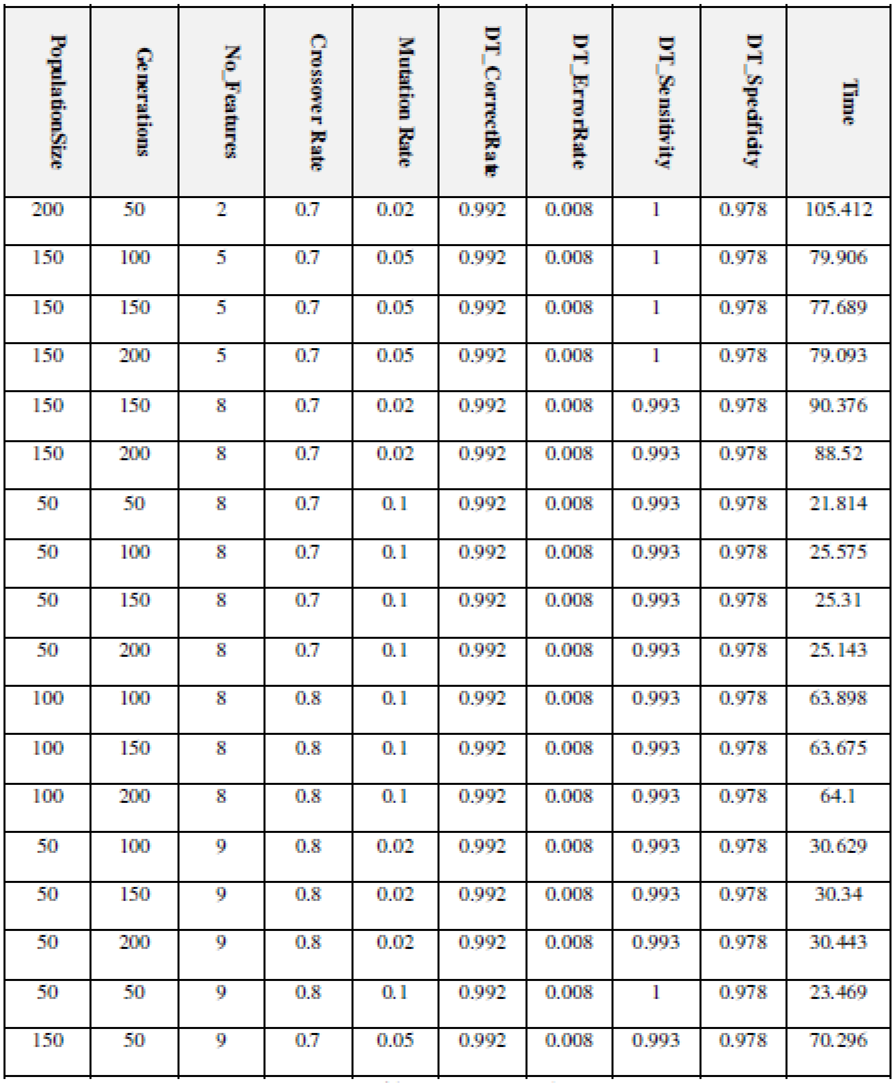
a)Create a Fitness Function for the Genetic Algorithm. In our case, the genetic algorithm tests small subsets of m/z values using the fitness function and then determines which m/z values to pass on to or remove from subsequent generations. The fitness function (biogafit) is passed to the genetic algorithm solver using a function handle. It maximizes the reparability of two classes using a linear combination of a posteriori probabilities and linear classifier error rates.
Fitness value = a posteriori probability + Error rate of a linear classifier

b) Set Genetic Algorithm Options. The GA function uses an options structure to store the algorithm parameters used to perform minimizations with the GAs. The gaoptimset function creates this options structure. The parameter values set for the GA are as follows:

- Population size: [50 100 150 200].
- Maximum number of generations: [50 100 150 200].
- Number of features: [1-10].
- Probability of crossover: [0.5 0.6 0.7 0.8].
- Probability of mutation: [0.02 0.05 0.1].
- @selectionroulette.
- @crossoversinglepoint.
- @mutationuniform.

c) Run GA to Find the Best Discriminative Features. We using the (ga) to start the GA function to decide the best feature values. We run the GA function with different times for all cases as a filter selection approach with DT. Then, we compute the DT correction rate (accuracy), the DT error rate, the DT sensitivity, and the DT specificity using 10-fold cross-validation. We also compute run time. Then, we compare the results to choice best accuracy trees. Table 3 lists the best GAs result along with their parameters. For example, in the first line we achieve 99.2 accuracy when using 200 population size, 50 generations, 0.7 crossover rate, 0.02 mutation Rate and only uses two features.The best results appears at (Table 3).

Table 3. Best GA results.



PopulationSize	Generations	No_Features	Crossover Rate	Mutation Rate	DT_CorrectRate	DT_ErrorRate	DT_Sensitivity	DT_Specificity	Time
200	50	2	0.7	0.02	0.992	0.008	1	0.978	105.412
150	100	5	0.7	0.05	0.992	0.008	1	0.978	79.906
150	150	5	0.7	0.05	0.992	0.008	1	0.978	77.689
150	200	5	0.7	0.05	0.992	0.008	1	0.978	79.093
150	150	8	0.7	0.02	0.992	0.008	0.993	0.978	90.376
150	200	8	0.7	0.02	0.992	0.008	0.993	0.978	88.52
50	50	8	0.7	0.1	0.992	0.008	0.993	0.978	21.814
50	100	8	0.7	0.1	0.992	0.008	0.993	0.978	25.575
50	150	8	0.7	0.1	0.992	0.008	0.993	0.978	25.31
50	200	8	0.7	0.1	0.992	0.008	0.993	0.978	25.143
100	100	8	0.8	0.1	0.992	0.008	0.993	0.978	63.898
100	150	8	0.8	0.1	0.992	0.008	0.993	0.978	63.675
100	200	8	0.8	0.1	0.992	0.008	0.993	0.978	64.1
50	100	9	0.8	0.02	0.992	0.008	0.993	0.978	30.629
50	150	9	0.8	0.02	0.992	0.008	0.993	0.978	30.34
50	200	9	0.8	0.02	0.992	0.008	0.993	0.978	30.443
50	50	9	0.8	0.1	0.992	0.008	1	0.978	23.469
150	50	9	0.7	0.05	0.992	0.008	0.993	0.978	70.296

4) Frequent Masses Analysis

Using the parameters in the previous table (Table 3), we obtain 42 different masses that give us the best accuracy results. Figure 6 shows the analysis of the frequencies with which the masses are selected, where mass 8073.585 and 244.3685 appear 10 times giving the best accuracy result.

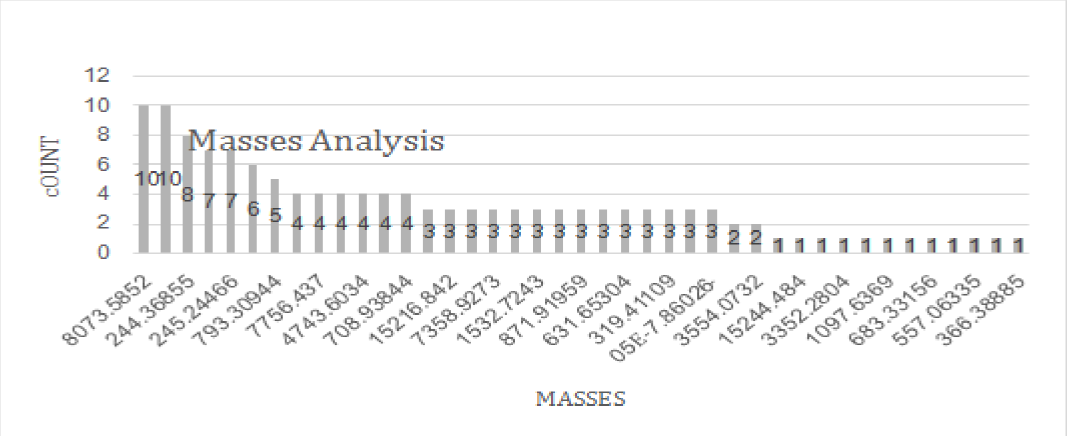


Figure 6. Masses analysis

5) Build multiple DTs From the Training Dataset Using Only Top MZ Values and Extract Frequent Rules.

This process involves built multiple DTs using different number of masses from the top frequency masses every times. Then, determine the most frequent rule in these trees for each number of masses. Steps below from A to J explain this process for each different number of masses from top two to top ten.For example**,** Using only the top two m/z values to build multiple DTs (x1=8073.5852 m/z and x2= 437.0239 m/z), which are the values extracted as top two features from the previous step. Then extract frequent rules from the multiple DTs that built by using the training dataset and compute the average coverage and accuracy of each rule using the test dataset. Note that we apply holdout validation 100 times, randomly reserve two-thirds of the dataset for training to build multiple DTs, and extract most frequent rule. The remaining one-third of the dataset is used for testing, the average coverage and accuracy are computed for the most frequent rules every time.

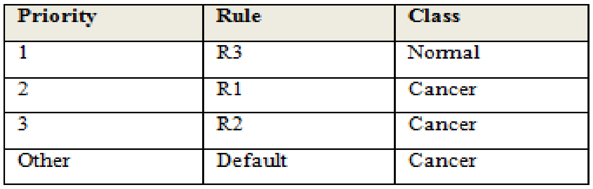
- R1: IF MZ (437.0239) >= 1.22269 THEN Class = Cancer.
- R2: IF MZ (437.0239) < 1.22269 and MZ (8073.5852) < 0.29102 THEN Class = Cancer.
- R3: IF MZ (437.0239) < 1.22269 and MZ (8073.5852) >= 0.29102 THEN Class = Normal.

Last, build a decision list using accuracy values.

Table 4.Rules accuracy values

Rule	**Average Coverage Percentage**	**Average Accuracy Percentage**
R1	100	95.74
R2	100	38.35
R3	100	97.62
Overall (R1+R2+R3)	100	98.80

Table 5.Decision list using accuracy values

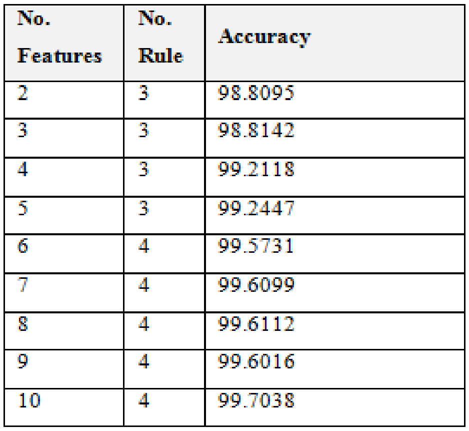


Priority	Rule	Class
1	R3	Normal
2	R1	Cancer
3	R2	Cancer
Other	Default	Cancer

6. DISCUSSION

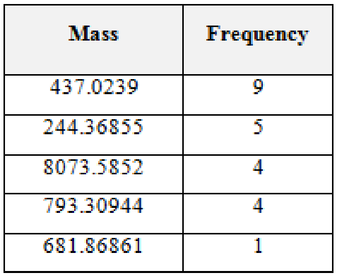
The results show that the GRC-MS classifier model achieves very good results when applied to the analysis of ovarian cancer datasets with different numbers of features or masses that used in the model (Table 6). We observed that 437.0239,244.36855, 8073.5852 and 793.30944 m/z were significantly discriminative masses that can be potential biomarkers for ovarian cancer. Table 7 lists the frequently occurring masses that play important roles in most of the rules.

Table 6. The GRC-MS classifier results.



No. Features	No. Rule	Accuracy
2	3	98.8095
3	3	98.8142
4	3	99.2118
5	3	99.2447
6	4	99.5731
7	4	99.6099
8	4	99.6112
9	4	99.6016
10	4	99.7038

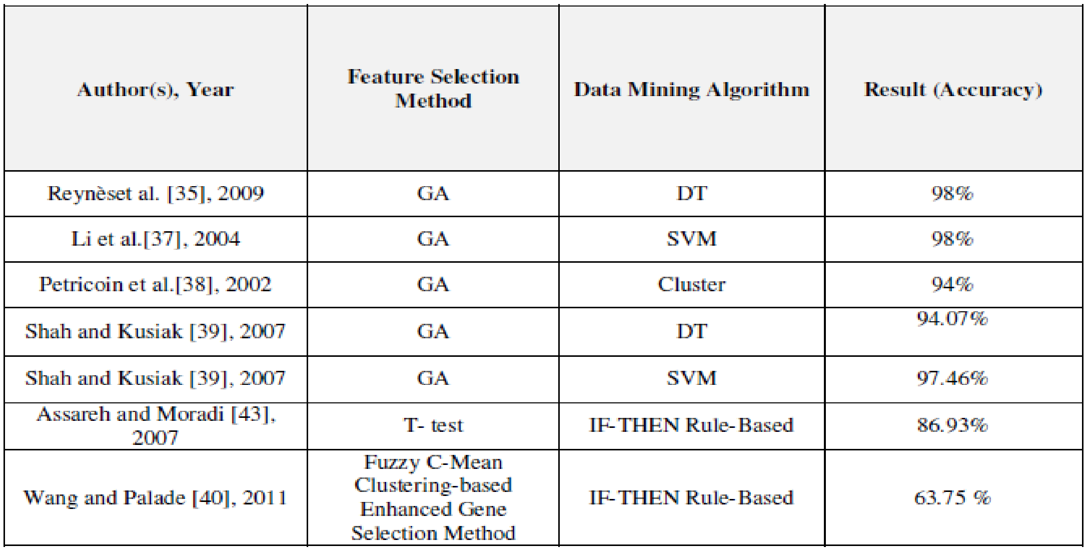
Table 7. Frequently occurring masses that play important roles in most of the rules



Mass	Frequency
437.0239	9
244.36855	5
8073.5852	4
793.30944	4
681.86861	1

In Table 8,shows that our GRC-MS classifier model provides highly competitive accuracy (99.7%) when compared to other existing classifier models, when applied to an ovarian cancer dataset. In addition, our model also provides highly comprehensible rules that facilitate the translation of raw data into easy-to-understand knowledge that can help experts.

Table 8. Results of some existing classifier models



Author(s), Year	Feature Selection Method	Data Mining Algorithm	Result (Accuracy)
Reynèset al. [35], 2009	GA	DT	98%
Li et al.[37], 2004	GA	SVM	98%
Petricoin et al.[38], 2002	GA	Cluster	94%
Shah and Kusiak [39], 2007	GA	DT	94.07%
Shah and Kusiak [39], 2007	GA	SVM	97.46%
Assareh and Moradi [43], 2007	T- test	IF-THEN Rule-Based	86.93%
Wang and Palade [40], 2011	Fuzzy C-Mean Clustering-based Enhanced Gene Selection Method	IF-THEN Rule-Based	63.75 %

7. CONCLUSION AND FUTURE WORK

Many studies have sought to increase the accuracy of diagnoses by analyzing MS data and finding biomarkers. Some of these studies have proposed approaches capable of high accuracy, sensitivity, and specificity, while other studies have failed to obtain satisfactory results. One major issue remains: How can an accurate model that avoids the “black box” limitation be built? The “black box” produces such problems as a lack of knowledge flow between the system and the expert. To address this problem and build a model capable of yielding accurate diagnoses that are easy for experts to understand, we used a ruled-based technique to build a classifier model to analyze MS data. Recently, significant attention has been paid to the use of rule-based classification techniques because of their unique ability to provide meaningful outcomes.

In addition, we apply a GA in the feature selection stage to increase the quality and accuracy of the p **GRC-MS** classifier model. In previous research, excellent results have been obtained through the combination of GA with different types of classifiers. In order to test the validity, accuracy, and performance of the **GRC-MS** model, we conducted an experimental study using open-source databases. In this experiment, we first applied several preprocessing steps to prepare the MS data for the **GRC-MS** model. These steps included reducing the noise in the data and the amount of data, identifying and extracting peaks, and normalizing and aligning the data. We found that the **GRC-MS** classifier model enhance the accuracy and meaningfulness of the MS data analysis results. As a future work, we aim to apply the **GRC-MS** model to another MS dataset or other high-dimension dataset, such as a microarray gene expression dataset. We also aim to develop more effective fitness functions for the GA.

REFERENCES

[1] H. Fernández, “Comparison of MALDI-TOF mass spectrometry data preprocessing techniques and their effect in sample classification.”

[2] M. Durairaj and V. Ranjani, “Data mining applications in healthcare sector a study,” International Journal of Scientific & Technology Research, vol. 2, pp. 29-35, 2013.

[3] Mass Spectrometry. (2015, Feb. 15). [Online]. Available: http://www.premierbiosoft.com/tech_notes/mass-spectrometry.html

[4] E. D. Hoffman and V. Stroobant, Mass Spectrometry: Principles and Applications. Belgium, John Wiley & Sons Ltd., 2007.

[5] P. Veltri, "Algorithms and tools for analysis and management of mass spectrometry data," Journal of Briefings in Bioinformatics, vol. 9, pp. 144-155, 2008.

[6] M. Cannataro, P. H. Guzzi, T. Mazza, G. Tradigo, and P. Veltri, "Preprocessing of mass spectrometry proteomics data on the grid," in Proc. 18th IEEE Symp. on Computer-Based Medical Systems, 2005, pp. 549-554.

[7] S. Bachmayer, "Preprocessing of mass spectrometry data in the field of proteomics," M.S. thesis, University of Helsinki, Finland, pp. 8-13, 2007.

[8] J. Zhang, E. Gonzalez, T. Hestilow, W. Haskins, and Y. Huang, "Review of peak detection algorithms in liquid-chromatography-mass spectrometry," Journal of Current Genomics, vol. 10, p. 388-401, 2009.

[9] R. Kandaa and R. Glendinning, "Mass spectrometry for environmental and wastewater monitoring," Journal of Spectroscopy Europe, vol. 23, pp. 15-27, 2011.

[10] M. Katajamaa and M. Orešič, "Data processing for mass spectrometry-based metabolomics," Journal of Chromatography A, vol. 1158, pp. 318-328, 2007.

[11] M. A. Rogers, P. Clarke, J. Noble, N. P. Munro, A. Paul, P. J. Selby, and R. E. Banks, "Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis identification of key issues affecting potential clinical utility," Journal of Cancer Research, vol. 63, pp. 6971-6983, 2003.

[12] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," Journal of Bioinformatics, vol. 23, pp. 2507-2517, 2007.

[13] D. Ward et al., "Identification of serum biomarkers for colon cancer by proteomic analysis," British Journal of Cancer, vol. 94, pp. 1898-1905, 2006.

[14] E. F. Petricoin et al., "Use of proteomic patterns in serum to identify ovarian cancer," The Lancet, vol. 359, pp. 572-577, 2002.

[15] G. Ge and G. W. Wong, "Classification of premalignant pancreatic cancer mass-spectrometry data using decision tree ensembles," Journal of BMC Bioinformatics, vol. 9, pp. 275-287, 2008.

[16] P. Yang and Z. Zhang, "A clustering based hybrid system for mass spectrometry data analysis," Journal of Pattern Recognition in Bioinformatics, pp. 98-109, 2008.

[17] C. Yang, Z. He, and W. Yu, "Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis," Journal of BMC Bioinformatics, vol. 10, pp. 4-14, 2009.

[18] "Smoothing," (2015, Feb. 22). [Online]. Available: http://www.wavemetrics.com/products/IGORPro/dataanalysis/signalprocessing/smoothing.htm.

[19] J. Prados, A. Kalousis, J. C. Sanchez, L. Allard, O. Carrette, and M. Hilario, "Mining mass spectra for diagnosis and biomarker discovery of cerebral accidents," Journal of Proteomics, vol. 4, pp. 2320-2332, 2004.

[20] A. Thomas, G. D. Tourassi, A. S. Elmaghraby, R. Valdes Jr, and S. A. Jortani, "Data mining in proteomic mass spectrometry," Journal of Clinical Proteomics, vol. 2, pp. 13-32, 2006.

[21] T. N. Vu and K. Laukens, "Getting your peaks in line: A review of alignment methods for NMR spectral data," Journal of Metabolites, vol. 3, pp. 259-276, 2013.

[22] Y. Su et al., "Diagnosis of gastric cancer using decision tree classification of mass spectral data," Journal of Cancer Science, vol. 98, pp. 37-43, 2007.

[23] I. Inza, P. Larrañaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains," Artificial Intelligence in Medicine, vol. 31, pp. 91-103, 2004.

[24] R. Blanco, P. Larranaga, I. Inza, and B. Sierra, "Gene selection for cancer classification using wrapper approaches," International Journal of Pattern Recognition and Artificial Intelligence, vol. 18, pp. 1373-1390, 2004.

[25] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," BMC Bioinformatics, vol. 7, p. 3, 2006.

[26] S. Ma and J. Huang, "Regularized ROC method for disease classification and biomarker selection with microarray data," Bioinformatics, vol. 21, pp. 4356-4362, 2005.

[27] F. Gullo, G. Ponti, A. Tagarelli, G. Tradigo, and P. Veltri, “A time series approach for clustering mass spectrometry data,” Journal of Computational Science, vol. 3, pp. 344-355, 2012.

[28] A. Vlahou, J. O. Schorge, B. W. Gregory, and R. L. Coleman, “Diagnosis of ovarian cancer using decision tree classification of mass spectral data,” Journal of Biomedicine and Biotechnology, vol. 5, pp. 391-404, 2003.

[29] Y. Su et al., “Diagnosis of gastric cancer using decision tree classification of mass spectral data,” Journal of Cancer Science, vol. 98, pp. 37-43, 2007.

[30] Y. D. Chen, S. Zheng, J.-K. Yu, and X. Hu, “Artificial neural networks analysis of surface-enhanced laser desorption/ionization mass spectra of serum protein pattern distinguishes colorectal cancer from healthy population,” Journal of Clinical Cancer Research, vol. 10, pp. 8380-8385, 2004.

[31] J. Zheng, “A comparison of naïve Bayes and logistic regression classifiers for mass spectrometry data,” Journal of Proteomics & Bioinformatics, 2007.

[32] M. Wagner, D. Naik, and A. Pothen, “Protocols for disease classification from mass spectrometry data,” Journal of Proteomics, vol. 3, pp. 1692-1698, 2003.

[33] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. San Francisco, CA, Morgan Kaufmann Publishers Inc., 2011.

[34] E. F. Petricoin et al., “Use of proteomic patterns in serum to identify ovarian cancer,” Journal of Mechanisms of Disease, vol. 359, pp. 572-577, 2002.

[35] C. Reynès, R. Sabatier, N. Molinari, and S. Lehmann, "A new genetic algorithm in proteomics: Feature selection for SELDI-TOF data," Computational Statistics & Data Analysis, vol. 52, pp. 4380-4394, 2008.

[36] M. S. Mohamad, S. Deris, S. Yatim, and M. Othman, "Feature selection method using genetic algorithm for the classification of small and high dimension data," in Proc. of the 1st Int. Symposium on Information and Communication Technology, 2004, pp. 13-16.

[37] L. Li et al., "Data mining techniques for cancer detection using serum proteomic profiling," Artificial Intelligence in Medicine, vol. 32, pp. 71-83, 2004.

[38] E. F. Petricoin et al., “Use of proteomic patterns in serum to identify ovarian cancer,” Journal of Mechanisms of Disease, vol. 359, pp. 572-577, 2002.

[39] S. Shah and A. Kusiak, "Cancer gene search with data-mining and genetic algorithms," Computers in Biology and Medicine, vol. 37, pp. 251-261, 2007.

[40] Z. Wang and V. Palade, "Building interpretable fuzzy models for high dimensional data analysis in cancer diagnosis," BMC Genomics, vol. 12, p. S5, 2011.

[41] H. W. Ressom et al., "Biomarker identification and rule extraction from mass spectral serum profiles," in proc IEEE Symp. Computational Intelligence and Bioinformatics and Computational Biology, 2006, pp. 1-7.

[42] J. M. Sorace and M. Zhan, "A data review and re-assessment of ovarian cancer serum proteomic profiling," BMC Bioinformatics, vol. 4, p. 24-35, 2003.

[43] A. Assareh and M. H. Moradi, "Extracting efficient fuzzy if-then rules from mass spectra of blood samples to early diagnosis of ovarian cancer," in Computational Intelligence and Bioinformatics and Computational Biology, 2007. CIBCB'07. IEEE Symposium on, 2007, pp. 502-506.

[44] Z. Wang and V. Palade, "Building interpretable fuzzy models for high dimensional data analysis in cancer diagnosis," BMC Genomics, vol. 12, p. S5, 2011.