Survey on different Methods for Classifying Gene Expression using Microarray Approach

Emad Mohamed Mashhour
Modern Academy
Computer Science Department

Enas M. F. El Houby
Systems & Information Department
Engineering Division
National Research Centre

Khaled Tawfik Wassif
Computer Science Department
Faculty of Computers and Information
Cairo University

Akram I. Salah
Computer Science Department
Faculty of Computers and Information
Cairo University

ABSTRACT

The recognizing and detecting process of genetic mutation becomes an important issue for research. There are various techniques that may help in detecting diseases, cancer and tumors. Microarrays are considered as type of representation for gene expression that may help in detection process. These gene expressions are used in analyzing samples that may be normal or affected, and help in diagnosis. To utilize the benefit of microarrays, machine learning algorithms and gene selection methods must be used to facilitate processing on microarrays and to overcome some challenges that may face microarrays. Challenges that may face Microarrays can be figured as a high dimensional data problem which is considered as an important challenges in different datasets. It suffers from redundant, irrelevant and noisy data. Solving this problem requires a method that simplifies this representation. Feature selection process can be a solution that may solve this important problem, through reducing the number of features to be used in clustering and classification. The problem can be defined as a selection of a small subset of genes from a set of gene expression data, recorded on DNA micro-arrays for classification. This survey observes some various techniques of classification, and gene selection methods such as filters and wrappers methods. To determine the suitable hybrid method or the powerful model that combine different techniques for detecting new or difficult mutated disease. And also introduces different emerging swarm intelligence techniques that prove its challenging ability in feature selection and classification in microarrays. These emerged techniques proved that there are upcoming approaches that can be used in detecting cancer. Swarm intelligence techniques proved that it can be hybridized with any mathematical or statistical techniques to gain better results.

Keywords

Microarray, Machine learning, Swarm intelligence, Feature selection, Classification

1. INTRODUCTION

Bioinformatics is ‘the collecting, archiving, organizing and interpreting biological data’. This goes beyond the collection and storage of data, to include the discussion of fundamental principles through classification, organization and interpretation. In other words it is considered as the application of computer technology to the management of biological information. It contains a lot of interesting challenges and areas that can help in solving more and more problems related to biological field [1].

The prognosis process is considered as an important process in controlling and preventing many diseases, specifically cancer, which will lead to discover new disease markers [1]. Detection of mutations or changes in gene expression patterns that occur as a result of the disease, will lead to accurate prediction of the reasons of cancer which will result in designing a powerful new therapies. Cancer is basically a disease “genes gone bad”. Many genes control the way cells grow, divide, and die. When these genes stop working and cell stop growing, this leads to tumor formation and cancer [1].

DNA microarray approach becomes an important factor for researchers who focus on understanding the growth and development of life through exploring the genetic causes of anomalies occurring in the human body. Microarray technology is a representation of genes that makes biologists be capable of monitoring expression of thousands of genes in a single experiment on a small chip [2].

Advances in microarray–based expression analysis research have led to the promise of cancer diagnosis using new molecular based approaches. Many studies and methodologies which analyze the gene expression have come up such as gene selection, classification and clustering.

Researchers try to do their best in discovering the suitable hybrid system in bioinformatics to detect accurately cancer and other diseases that may lead to gene mutation. Cancer is an abnormal cell-growth occurring in human body and may originate from any of the areas or organs.

The disorder in certain cells in human body can be very dangerous, or even fatal, if ignored for long. It can be developed to tumors that may spread. Such tumors spread to various parts of the body via bloodstream. Therefore Research requires deep study of most common Cancer in men and women including Lung, Prostate, Breast, Oral Cancer and their recognition.

2. BACKGROUND

2.1 DNA Microarray Technology

Microarrays [3] are a well-established technology to show the expression of many genes in a single reaction whose applications range from cancer diagnosis to drug response. The expression level of genes is given by the amount of mRNA bounding to each entry. The aim is to find either set of genes that can deeply represent a particular disease states, experimental condition and highly correlated genes that share common biological features.

Scientists used DNA microarrays to specify the expression levels of huge numbers of genes concurrently. Significant information can be extracted from these genes by using machine learning techniques.

The usefulness of DNA Microarray technology can be listed as

1. Can follow the activity of many genes at the same time.
2. Can get a lot of results fast.
3. Can compare the activity of many genes in diseased and healthy cells.
4. Can classify diseases into subgroups.

In fact, Microarray analysis has proved that accurate cancer diagnosis can be achieved by performing microarray data classification, which is done by constructing classifiers to compare the gene expression profile of unknown cancer status to a stored gene expression profile from tissues of known cancer status. Microarray approach can lead us to the knowledge of how each gene can be helpful, informative or noisy. Microarrays measures gene expressions of thousands of genes in parallel. This data is expressed in an array that contains huge data of genes called (huge dimensionality data). This data expression structure full of active relevant genes and inactive irrelevant genes. Disadvantage of this representation is that it can be a reason to misleading of classification process because of its noisy and irrelevant data. All cancer datasets that results from microarrays may contain number of genes that exceed number of samples, which is called high dimensionality problem [4]. Therefore there must be a feature selection method that may help in reducing the size of the feature set (gene set). Microarray structure is shown in figure [1].

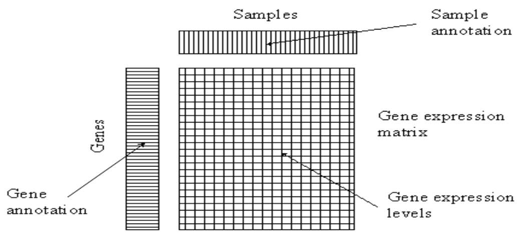


Fig 1: Microarray Structure

2.2 Diagnosis process of Disease Using Microarray

Usually, microarray diagnosis process contains two successive steps, feature selection and classification. So far, many machine learning algorithms have been introduced, and many of them have been employed for detecting mutations, including the techniques of feature selection and classification, for example K-NN, support vector machines, neural networks ,and the emerging field of swarm intelligence algorithms that were successfully used in these two steps. Most of researches attempt to choose an optimal subset of genes and then build a classification model based on these selected genes.

2.2.1 Feature Selection Techniques

Microarray data often contains many irrelevant and redundant features, which affect the speed and accuracy of most learning algorithms. Therefore, feature selection is widely used to addresses this problem in microarray data analysis; in biological field specifically gene processing feature selection is known as gene selection.

In the context of classification, feature selection techniques can be divided into three types, according to the classifier needs: filter methods, wrapper methods, and embedded methods as shown below in table (1) [4].

Table 1: Different methods of feature selection

Filter techniques	Filter techniques evaluate how much the features are relative to the problem by checking properties of the data. In most cases a feature relevance score is calculated, and the low scoring features are removed. Afterwards, high ranked features are presented as input to the classification algorithm.
Wrapper techniques	Wrapper techniques evaluate subsets of features according to how informative they are to a given predictor. The method conducts a search for a good subset using the learning algorithm itself as part of the evaluation function. The problem boils down to a problem of stochastic state space search.
Embedded techniques	Embedded techniques perform feature selection as part of the learning procedure and are usually specific to given learning machines. Examples are classification trees, random forests, and methods based on regularization techniques.

The processing of microarray data expression can be structured as a set of sequential steps, as data feeds along next steps; it becomes more and more refined. The goals of such pipeline are: (i) to identify and remove the noise due to the experimental procedure, (ii) to extract the real value of expression for each gene, (iii) to match each probe with the corresponding nucleotide sequence, and, (iv) to enrich such information by using functional annotations. Each step can be performed by using different algorithms. A typical workflow for analyzing microarray data is structured on four main phases: (i) summarization and normalization, (ii) annotation, (iii) statistical or data mining analysis, and (iv) biological interpretation [5].

1. **Normalization**: is the process of reducing unwanted variation either within or between arrays. Typical assumptions of most major normalization methods are (one or both of the following):[6]
   - Only a minority of genes are expected to be differentially expressed between conditions.
   - Any differential expression is as likely to be up-regulation as down-regulation.
2. **Summarization:** Reducing multiple measurements on the same gene down to a single measurement by combining in some manner [6].
3. **Annotation:** Probably, the most important thing you want to know is what the genes or their products are concerned with, i.e. their function**.**
4. **Statistical Analysis:** From a statistical point of view, the number of genes is sometimes larger than the number of samples, which can cause the process of diagnosis to miss classification. Selecting features/genes is considered

as a process of selecting the most informative genes that may improve the prediction performance of the classification process, and to avoid the problem of dimensionality. Another improvement for this approach is to reduce the computation time and understand the biological process outcomes of these kinds of gene expression data [6].

5. **Biological Interpretation:** Interpreting the huge data generated by microarray technology poses a number of unusual new challenges. To be able to interpret microarray data, you must have a sufficient number of replicate measurements. Such replication is necessary to evaluate which results will have real predictive value. Interpretation of microarrays needs dimensionality reduction. This means that the factors or phenomena that appears in the gene expression data need to be extracted and emphasized. There are two principal ways of achieving this. The first one, cluster analysis, is exclusively data driven, the second one, annotation analysis, is knowledge-driven [7].

2.2.2 Microarray Classification

In microarray classification tasks, samples belong to one of several classes such as cancer dataset or normal dataset, the goal is to classify these samples and produce classified samples based on its microarray measurements. Training classifiers on such small-sample high-dimensional data sets is a challenging problem that has received an increasing attention from the research community. A standard way of addressing the challenge is to perform feature selection as a pre-processing step and to follow it by applying a classification algorithm that controls model complexity through regularization [8].

Machine learning is a scientific approach that addresses some questions such as, how can systems be programmed to automatically learn and to improve with experience. Learning in this context is not considered a real learning process but recognizing complex patterns and make intelligence decisions based on data. Machine learning develops algorithms that discover knowledge from specific data and experience, based on computational principles.

2.2.2.1 Non Swarm Intelligence techniques

Classification refers to a procedure that assigns data objects to a set of classes. Some well-known classification techniques that are used in data mining and many other fields such as: [9]

- **Decision Tree**

A decision tree is a tree-structured classifier. It is a method that learns decision tree using a recursive tree growing process. Each test corresponding to an attribute is evaluated on the training data using a test criteria function. The test criteria function assigns each test a score based on how well it partitions the data set, the highest score is selected and placed at the root of the tree. The sub-trees of each node are then grown recursively by applying the same algorithm to the examples in each leaf. The algorithm terminates when the current node contains either all positive or all negative examples [9].

- **Neural Network**

An Artificial Neural Network (ANN), or commonly just called neural network (NN), is an artificial intelligence technique that simulates the operations and processes of the human brain. It is represented as a set of neurons which is interconnected to each other. The main feature of this technique is learning/training data, and classification of data. Structure and weight of any neural network changes according to information in input [10].

- **Support Vector Machine**

SVMs are a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors in an n-dimensional space, SVM will construct a separating hyper plane in that space, which maximizes the margin between the two data sets. To calculate the margin, two parallel hyper planes are constructed, one on each side of the separating hyper plane [11].

2.2.2.2 Swarm Intelligence Techniques

Swarm intelligence system consists typically of a population of simple agents interacting locally with one another and with their environment. The inspiration often comes from nature, especially biological systems. The agents follow very simple rules, and although there is no centralized control structure dictating how individual agents should behave. Interactions between agents may lead to the emergence of "intelligent" global behavior. Swarm intelligence algorithm such as Ant Colony Optimization (ACO), Artificial Bee Colony optimization (ABC), Particle Swarm optimization (PSO) and other systems proves its powerful ability in biological problems. Swarm intelligence is quite a general concept that multiple agents interact and exchange information [12].

- **Artificial Bee Colony**

The main idea behind artificial bee colony is to simulate the foraging behavior of the bees such as exploration, exploitation, recruitment and abandonment. ABC algorithm is considered as an intelligence swarm, due to its foraging behavior that may assist in the clustering, classification and any optimization methods.

ABC algorithm contains artificial group of bees that consist of three types of bees: employed bees, onlooker bees and scout bees. In ABC, each food source position is considered as a possible solution and the nectar amount of food source is considered as the quality (fitness) of the solution [13].

- **Bat Algorithm**

Bat algorithm is emerging algorithm in swarm intelligence algorithm, which proves its capability in many problems. Bats have an advanced capability called echolocation. Echolocation process works as a type of sonar that let the bat emit some loud and short pulses of sound in the space, and then wait until this pulse hit anything, this pulse returns to the bat again. This returned pulse can let the bat compute the distance and how far this obstacle. This capability let the bats can distinguish between an obstacle and preys. This capability can be considered as a classification process. Bats usually fly randomly to search for prey. According to that, some parameters must be considered with the bats' foraging behavior, such as velocity, position, fixed frequency (fmin), varying wavelength and loudness [14].

- **Ant Colony Optimization**

Ant colony optimization ACO is an algorithm that simulates the ant behavior in nature. ACO algorithm takes its mechanism from the foraging behavior of ant. Ants lay pheromone on the ground where it walks. This pheromone helps ants to determine a path where other ants can follow. Ants tend to select the paths that contain higher concentration of pheromone. Through this methodology ants are capable of transporting food in an effective way. ACO help in searching for food in a faster way. It can be used in any optimization method. Artificial ants generate artificial pheromone to update the path; this path will describe to next artificial ant the new path it will choose [15].

- **Particle Swarm Optimization**

It is considered as a population-based method, this algorithm simulate the behavior of the bird flocking. Each bird represents solution in the search space. Each solution is considered as a particle, which has a fitness value that can be evaluated with. The performance of particle comes from sharing each particle a valuable information, in which each particle mimic the whole behavior of all the swarm, which will help in optimizing the objective of the swarm. This behavior can definitely help in search problems for best solutions [16].

3. EVALUATION METHODOLOGY

In classifying unseen data, depending on the class predicted by the classifier and the true class of the patient (Control or infected), four possible types of results can be observed for the prediction as follows:

1. True positive—the result of the patient has been predicted as positive (Cancer) and the patient has cancer.
2. False positive—the result of the patient has been predicted as positive (Cancer) but the patient does not have cancer.
3. True negative—the result of the patient has been predicted as negative (Control), and indeed, the patient does not have cancer.
4. False negative—the result of the patient has been predicted as negative (Control) but the patient has cancer.

Let TP, FP, TN, and FN, respectively, denote the number of true positives, false positives, true negatives, and false negatives. For each learning and evaluation experiment, Accuracy, Sensitivity, and Specificity defined below are used as the fitness or performance indicators of the classification:

Accuracy= (TP+TN)/ (TP+TN+FP+FN).

Sensitivity=TP/ (TP+FN).

Specificity=TN/ (TN+FP).

4. RELATED WORK

There exists considerable literature on microarray applications. To explore some of the hybridization systems available for this technology, a brief survey will be presented in order to show how powerful to hybridize microarray with machine learning techniques. Most algorithms in machine learning are divided into, non-swarm intelligence based algorithms and swarm intelligence algorithms. Researchers were concerned in their work on two important aspects:

1. To identify molecular signatures associated with known classes.
2. To discover new classes.

After detecting and filtering gene expression datasets, it is often necessary to accurately classify samples into known groups according to the features of the gene expression. There are a quite few methods to classify samples for instance support vector machines (SVM), (PAM) prediction analysis of microarrays, classification and regression trees (CART), k-Nearest-Neighbor (k-NN) and others.

4.1 Applied non Swarm Intelligence algorithms in Microarrays

Chaunliangchen, et al [17] have used technique called AIRS (artificial immune recognition system) in order to perform microarray data (cancer, disease or normal tissues) classification. Three other classifiers have been used for comparing results; in this technique memory cell has been used for training samples in order to build a classifier. The experiment is applied on some diseases represented as data set such as colon cancer, brain tumor, and nine tumors. After experiments, results reveal that AIRS gives more higher results as a classifier from the other three machine learning (KNN, OneR, Naïve Bayes)

Nazario D.Ramirez Beltran, et al [18], in their research they focused on classifying human tumors based on microarray information. They used an algorithm called projection algorithm. It depends on the geometrical projection principle. It relies on a simple theory of detection and recognition through training samples and then comparing it with new samples. It works as follows; the dataset of training has been represented through two vectors: cancer vector and normal vector.

$$\theta_i=\cos^{-1}\left(\frac{hi \cdot ci}{\|hi\|\|ci\|}\right), \quad i=1, 2, \ldots\ldots, N$$

The tissue selected from the validation data has been classified either cancerous or normal. This classified tissue has been used to create third vector projected on the other two vectors.

The projection angle used in this algorithm is the angle between the normal vector and the cancer vector. The inner product of the two vectors is used to calculate the angle. Genes with the largest angle are selected as genes that have largest expression. Projection angle technique proved its success over Neural Networks, Fisher Discriminant, and Logistic Regression in detecting cancer tissue.

A. Bharathi, A.M.Natarajan [19] presented a technique for finding the smallest set of genes that can ensure accurate classification of cancer from microarray data. A type of supervised machine learning algorithm called (ANOVA) stands for analysis of variance has been used in this research. Researchers need to rely on the smallest set of genes to get the higher accuracy in classification. First they compute the ranking of gene using ANOVA. ANOVA is a technique for analyzing experimental data under various conditions. Then SVM is used as a classifier. The technique is then compared with other classifiers such as T-test. The results obtained were for the hybridization techniques of ANOVA and SVM showed that the proposed approach classifies accurately with the minimum number of genes. Researchers obtain good results in classification through a combination of two genes from the selected genes.

Xiaosheng wang, Osamu gotoh [20] presented a technique called α depended degree –based feature selection. Researchers aim to solve the problem of the imbalance between the feature numbers and instance numbers in microarray data based gene expression. Researchers performed gene selection using the α depended degree criterion, which is done by tuning the value of α. The accurate classification with a small size of genes is better than with large number. Researchers applied their techniques on Colon Tumor, (Central Nervous System) Tumor, (Diffuse Large B-Cell Lymphoma), Leukemia 1, AML, Lung Cancer, Prostate

Cancer, Breast Cancer, and Leukemia The results has been compared with various techniques such as NB (Naïve Bayes), DT (Decision Tree), SVM (Support Vector Machine) and kNN (k-nearest neighbor). The results reveal that k-NN classifier had better performance under seven α values.

Katerina N. Karayianni, et al [21] considered that performing clustering in microarrays is a challenging process due to the nature of the datasets used which contains noise. In this work fuzzy clustering method is used with viewpoints. The target of that work is to perform or construct a prediction model that aid to the identification of unlabeled samples. The viewpoints used in the fuzzy clustering with viewpoints method have been constructed by computing the average expression value for every feature (probe/gene) among the samples that have the particular label. To guide the clustering process, the previously available microarray expression data was used and introduced as viewpoints in the clustering process. The technique was applied on breast cancer, brain cancer, AML and MLL datasets. The results show a quite positive impact of using viewpoints to improve the clustering process. The method was compared to the clustering algorithms of k-means, fuzzy c-means, affinity propagation, as well as a method of clustering microarray data that is based on prior biological knowledge, and has shown comparable and improved results over them.

Sudip Mandal and Indrojit Banerjee [22] applied artificial neural network to detect cancer, and to help physicians to diagnosis accurately. Artificial intelligence techniques provide solutions to real problems such as analyzing microarrays, and detecting cancer features from it. Researchers used a special kind of ANN called multilayer feed forward neural network (MLFF). Performance of ANN always depends on the design of the neural network such as number of hidden layers, number of nodes and weights. Different datasets was tested in their research such as breast cancer with different status and the lung dataset. Two ways are used for analysis cross validation and new case test dataset. Datasets was divided into 80% for training and 20% for testing. Different datasets were used such as breast cancer and its different status and lung dataset. Due to the noisy dataset in microarray, the accuracy was 96% after cross validation and 94% for testing new dataset. ANN was designed with single hidden layer, but it was clear that the structure of the ANN can be tuned, in other words number of hidden layers and nodes can be increased or decreased. The researchers try to increase both the hidden layers and nodes until certain level, so they can get better result. Increasing these values above certain level may cause the results to have low accuracy. Results of their tuning experiment are shown in figure (2) below. Three hidden layers and three nodes was the best structure to the proposed ANN to give the best result for classification.

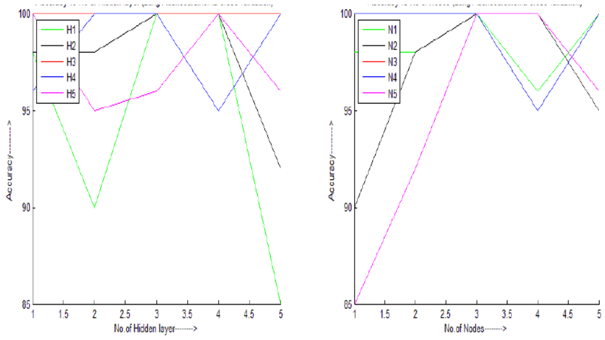


Fig 2: High Percentage at Three Hidden Layer and Three Nodes [22]

4.2 Applied Swarm Intelligence algorithms in Microarrays

Hualong Yu, et al [23] detects that there may be imbalance problem that may occur in microarray because of the huge, noisy and irrelevant samples. Due to this problem researchers intended to use swarm intelligence techniques in order to solve this problem. Researchers used technique called ACOSampling inspired by ACO (ant colony optimization) algorithm in order to eliminate noisy and irrelevant features (genes) under a process called feature selection. They applied their technique on four benchmark microarray datasets, including Colon dataset, CNS (Central Neural System) dataset, Lung cancer dataset and Glioma dataset. Datasets are divided into two parts, training subset and validation subset in order to test the data with a certain classifier. Researchers used datasets with binary classes and multi class. SVM technique is used as a classifier due to its performance in high dimensional data with small number of samples. Performances varies due to the cross validation, each dataset has a suitable number of features that maybe used in classification in order to yield a better performance.

Mustafa Serter Uzer, et al [24] used a hybrid approach that merges between SVM as a classifier and ABC (artificial bee colony) algorithm as an attribute reduction tool. Researchers use medical datasets which has the same characteristics as microarrays, high dimensional data. The experiment was applied on hepatitis, liver disorder disease and diabetes datasets. Researchers intended to use algorithm ABC in clustering as a selection criteria in order to reduce the search space of features. Main objective is to serve the feature selection phase. Then use the SVM method as a classifier as shown in figure (3) and k-fold cross validation in order to improve the classification performance.

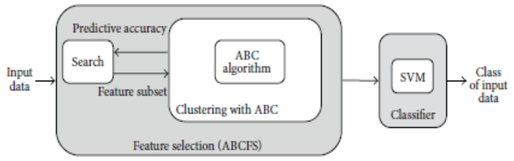


Fig 3: Proposed Model [24]

But one of the points that can be seen is that performance of classification relies sometimes on the kind of dataset. But in their comparative study they get the highest result with 10 cross validation.

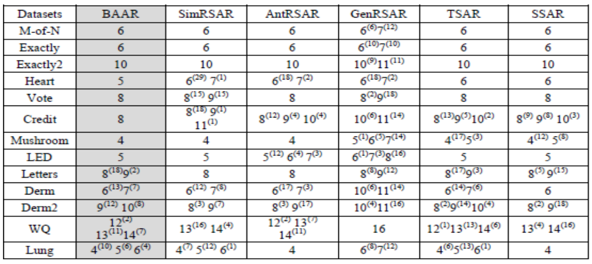
Hesham Arafat, et al [25] provided a method to find a way to reduce features, to select the optimal subset of features that may improve the classification process. Researchers used rough set theory (RST) to evaluate how much the selected subset of features is informative. This process is done using an improved searching process of features called ant colony optimization (ACO). The aim of their proposed model is to find the most reliable subset that may be used effectively in prediction (classification). RST is useful in dividing space into clusters and thus it can help in reducing attributes and avoiding redundant attributes. Researchers applied the proposed algorithm on breast cancer and other datasets. The proposed model gives the best result in reducing attributes as shown in table (2).

Table 2: The reduces of the attributes [25]

Data set	Inst	Feat	MinRedu	RSQR	RSFSACO
Audiology	200	70	-	20	13
Breast-cancer	699	10	4	4	4
Mushroom	8124	23	4	6	4
Wine	178	14	5	6	5
Vote	435	17	9	12	9

AHMED MAJID TAHA, ALICIA Y.C. TANG [26] tried to explore the problem of attribute reduction. This is considered one of the main problems in medical, microarray and other datasets. This suffers from irrelevant and redundant of features. Researchers proposed an optimization method to help in feature selection relies on swarm intelligence algorithms. Bat algorithm has been implemented in order to reduce attributes. Researchers use RST due to its performance in supporting approximations in decision making. BAT algorithm is merged with RST due to its capability in searching for prey or in other meaning in finding informative attributes (genes) in medical datasets. Bat algorithm for attribute reduction (BAAR) algorithm relies on the concept of echolocation of micro bats. Rough set is calculated and its parameters are used in the BAAR algorithm. The proposed technique is applied on Heart disease, lung cancer dataset and other datasets. By comparing results researchers claim that their proposed model give higher performance in feature selection than some known methods and the same performance compared with other methods as shown in table (3).

Table 3: number of most informative genes selected by BAAR [26]

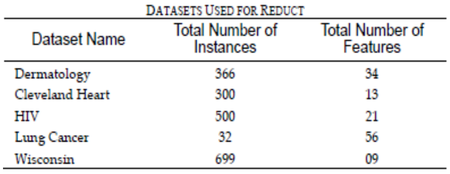


Datasets	BAAR	SimRSAR	AntRSAR	GenRSAR	TSAR	SSAR
M-of-N	6	6	6	$6^{(6)}7^{(12)}$	6	6
Exactly	6	6	6	$6^{(10)}7^{(10)}$	6	6
Exactly2	10	10	10	$10^{(9)}11^{(11)}$	10	10
Heart	5	$6^{(29)}\ 7^{(1)}$	$6^{(18)}\ 7^{(2)}$	$6^{(18)}7^{(2)}$	6	6
Vote	8	$8^{(15)}\ 9^{(15)}$	8	$8^{(2)}9^{(18)}$	8	8
Credit	8	$8^{(18)}\ 9^{(1)}\ 11^{(1)}$	$8^{(12)}\ 9^{(4)}\ 10^{(4)}$	$10^{(6)}11^{(14)}$	$8^{(13)}9^{(5)}10^{(2)}$	$8^{(9)}\ 9^{(8)}\ 10^{(3)}$
Mushroom	4	4	4	$5^{(1)}6^{(5)}7^{(14)}$	$4^{(17)}5^{(3)}$	$4^{(12)}\ 5^{(8)}$
LED	5	5	$5^{(12)}\ 6^{(4)}\ 7^{(3)}$	$6^{(1)}7^{(3)}8^{(16)}$	5	5
Letters	$8^{(18)}9^{(2)}$	8	8	$8^{(8)}9^{(12)}$	$8^{(17)}9^{(3)}$	$8^{(5)}\ 9^{(15)}$
Derm	$6^{(13)}7^{(7)}$	$6^{(12)}\ 7^{(8)}$	$6^{(17)}\ 7^{(3)}$	$10^{(6)}11^{(14)}$	$6^{(14)}7^{(6)}$	6
Derm2	$9^{(12)}\ 10^{(8)}$	$8^{(3)}\ 9^{(7)}$	$8^{(3)}\ 9^{(17)}$	$10^{(4)}11^{(16)}$	$8^{(2)}9^{(14)}10^{(4)}$	$8^{(2)}\ 9^{(18)}$
WQ	$12^{(2)}\ 13^{(11)}14^{(7)}$	$13^{(16)}\ 14^{(4)}$	$12^{(2)}\ 13^{(7)}\ 14^{(11)}$	16	$12^{(1)}13^{(13)}14^{(6)}$	$13^{(4)}\ 14^{(16)}$
Lung	$4^{(10)}\ 5^{(6)}\ 6^{(4)}$	$4^{(7)}\ 5^{(12)}\ 6^{(1)}$	4	$6^{(8)}7^{(12)}$	$4^{(6)}5^{(13)}6^{(1)}$	4

P. Ganesh Kumar, et al [27] proposed a method that can analyze datasets and extract information (informative attributes) that helps in diagnosis. Fuzzy expert system is targeted due to its results that are understandable and interpreted diagnostic information. And since data in microarray datasets or other datasets can't be reliable information due to the irrelevant and duplicated information, which means uncertainty information. Fuzzy system can deal with uncertainty information. Researchers rely in building this method on two swarm intelligence algorithms, ant colony optimization (ACO) and artificial bee colony (ABC). The aim of this hybrid method is to create if-then rules that may help with a membership function in choosing the suitable attributes. Researchers used ant colony optimization (ACO) to search for the best simple set of rules, in which artificial bee colony are used to tune the membership function. After extracting the set of rules, it is evaluated through the membership or objective function. Researchers applied their model on diabetes datasets and claim that their model creates simple set of rules that can yield the best reduction for attribute which can affect the classification accuracy.

N. Suguna, Dr. K. Thanushkodi [28] proposed a method for searching for the optimal subset of features after dataset reduction. Researchers used the medical domain datasets as shown in table (4) in testing their model and comparing their method with other methods.

Table 4: The datasets tested before reduction [28]

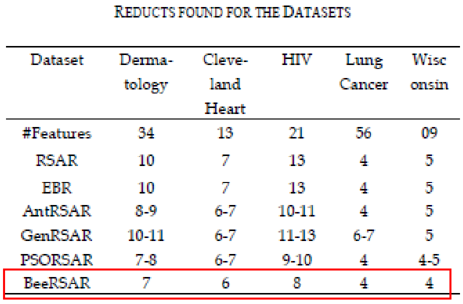


DATASETS USED FOR REDUCT

Dataset Name	Total Number of Instances	Total Number of Features
Dermatology	366	34
Cleveland Heart	300	13
HIV	500	21
Lung Cancer	32	56
Wisconsin	699	09

Researchers intended to use swarm intelligence techniques hybrid with rough set theory. Researchers used rough set attribute reduction (RSAR) in feature selection inspired by RST, and hybrid this method with bee colony optimization (BCO). They compared between different other hybrid techniques such as ant colony with rough set (Ant-RSAR), particle swarm optimization with rough set (PSO-RSAR) and genetic algorithm with rough set (Gen-RSAR).

Table 5: BeeRSAR performance in reducing features [28]



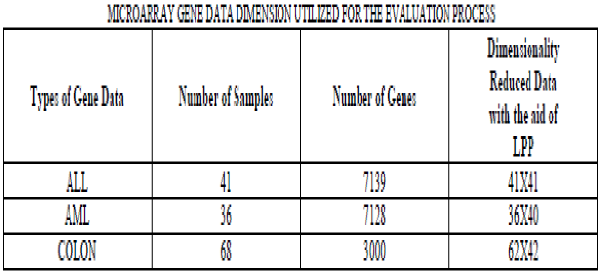
REDUCTS FOUND FOR THE DATASETS

Dataset	Dermatology	Cleveland Heart	HIV	Lung Cancer	Wisconsin
#Features	34	13	21	56	09
RSAR	10	7	13	4	5
EBR	10	7	13	4	5
AntRSAR	8-9	6-7	10-11	4	5
GenRSAR	10-11	6-7	11-13	6-7	5
PSORSAR	7-8	6-7	9-10	4	4-5
BeeRSAR	7	6	8	4	4

Researchers claim that. Bee colony with rough set (Bee-RSAR) performs well in reducing attributes as shown in table (5), but it takes more time in reducing attributes.

K.Sathishkumar, et al [29] proposed a new method using cluster analysis approach. The idea is to use an algorithm that can reduce and solve the problem of high dimensionality. Therefore another method was used to cluster samples according to their phenotype or pattern. Locality sensitive discriminant analysis (LSDA) is an algorithm that can be used as a dimensionality reduction method. The method of reduction was applied on three cancer dataset and it showed the effect of reduction as shown in table (6).

Table 6: The dimensionality reduction [29]



MICROARRAY GENE DATA DIMENSION UTILIZED FOR THE EVALUATION PROCESS

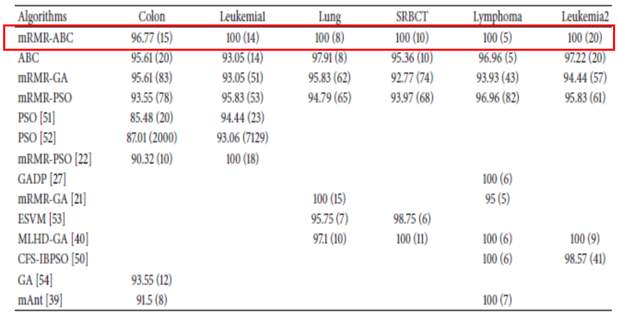
Types of Gene Data	Number of Samples	Number of Genes	Dimensionality Reduced Data with the aid of LPP
ALL	41	7139	41X41
AML	36	7128	36X40
COLON	68	3000	62X42

After the reduction phase, researchers intended to use a swarm intelligence technique in clustering based on Fuzzy C-means (FCM) in order to cluster gene expression according to certain features or pattern. Artificial bee colony is used hybrid with Fuzzy C-means (FCM). Algorithm called MoABC based FCM is used to cluster microarray gene expression of a human, taking the advantage of the bees in searching for honey with simplicity and robustness and FCM with distance feature between the measured data and the clustered data. The results showed that the proposed model gives superior result in accuracy, correlation, distance and error rate.

Rodrigo Yuji Mizobe Nakamura, et al [30] used an algorithm called bat algorithm. Researchers used this algorithm to select features that may be used in classification. They used a modified bat algorithm called binary bat algorithm (BBA) as a feature selection method. It acts by giving each bat a binary coordinate that may show whether the feature will be selected or not. It searches in a space to a position. These positions are represented in binary values which represent the presence or absence of the feature. Researchers used the k-fold cross validation, dataset is divided into 70% and 30% subsets, that can be used in training and validating the classifier. Optimum path forest classifier is used as a classifier in order to test the features selected. Optimum path forest act as a supervised learning algorithm that depends on pre-knowledge from a given features. Results showed that their method yield performance same like firefly and particle swarm optimization and better than gravitational search algorithm and harmony search.

Hala Alshamlan, et al [31] tried to create a profile for microarray gene expression. Researchers applied feature selection phase using an innovation of a hybrid system between swarm intelligence algorithm called (ABC) artificial bee colony and a method called (mRMR) minimum redundancy maximum relevance in order to select an informative features (genes) for better prediction for classification. mRMR is considered as a method to reduce noise and remove irrelevant features, as well as reducing dimensionality. ABC goes under a mission to estimate the best predictive genes. After extracting the best features, SVM is used as a classifier. Features generated from ABC are used as a subset ready to train SVM classifier. Cancer datasets have been used in testing.

Table 7: Performance of the model in [30] compared with other technique [31]

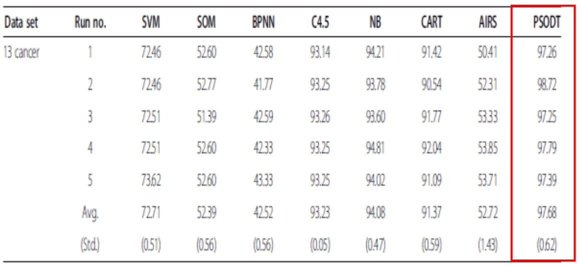


Algorithms	Colon	Leukemia1	Lung	SRBCT	Lymphoma	Leukemia2
mRMR-ABC	96.77 (15)	100 (14)	100 (8)	100 (10)	100 (5)	100 (20)
ABC	95.61 (20)	93.05 (14)	97.91 (8)	95.36 (10)	96.96 (5)	97.22 (20)
mRMR-GA	95.61 (83)	93.05 (51)	95.83 (62)	92.77 (74)	93.93 (43)	94.44 (57)
mRMR-PSO	93.55 (78)	95.83 (53)	94.79 (65)	93.97 (68)	96.96 (82)	95.83 (61)
PSO [51]	85.48 (20)	94.44 (23)				
PSO [52]	87.01 (2000)	93.06 (7129)				
mRMR-PSO [22]	90.32 (10)	100 (18)				
GADP [27]					100 (6)	
mRMR-GA [21]			100 (15)		95 (5)	
ESVM [53]			95.75 (7)	98.75 (6)		
MLHD-GA [40]			97.1 (10)	100 (11)	100 (6)	100 (9)
CFS-IBPSO [50]					100 (6)	98.57 (41)
GA [54]	93.55 (12)					
mAnt [39]	91.5 (8)				100 (7)	

Researchers compared their methods with other hybrid systems such as minimum redundancy maximum relevance with genetic algorithm (mRMR-GA), minimum redundancy maximum relevance with particle swarm optimization (mRMR-PSO), and other methods; the proposed model gives superior results than other methods as shown in table 7.

Kun-Huang Chen, et al [32] proposed a solution for the problem of extracting the important and informative genes in order to use in classification for cancerous datasets. Accuracy in selecting informative genes may lead to the main reason of the occurrence of cancer. Researchers used a hybrid system that merges DT (decision tree) as a classifier and PSO (particle swarm optimization) as a gene selection method. Their method was compared with other techniques such as SVM, self-organizing map (SOM), C4.5, back propagation neural network (BPNN), naïve Bayes (NB) and artificial immune recognition system (AIRS). Researchers proved that their technique may give better results through testing their technique on different cancer dataset. In order to test different datasets, researchers aim to use cross validation technique. Fivefold cross validation was applied to ensure the reliability of the results. PSODT showed that it can earn better performances than other techniques applied on the datasets. Accuracy of techniques is shown in table (8) below.

Table 8: Classification accuracy of techniques against PSODT (%) [32]



Data set	Run no.	SVM	SOM	BPNN	C4.5	NB	CART	AIRS	PSODT
13 cancer	1	72.46	52.60	42.58	93.14	94.21	91.42	50.41	97.26
	2	72.46	52.77	41.77	93.25	93.78	90.54	52.31	98.72
	3	72.51	51.39	42.59	93.26	93.60	91.77	53.33	97.25
	4	72.51	52.60	42.33	93.25	94.81	92.04	53.85	97.79
	5	73.62	52.60	43.33	93.25	94.02	91.09	53.71	97.39
	Avg.	72.71	52.39	42.52	93.23	94.08	91.37	52.72	97.68
	(Std.)	(0.51)	(0.56)	(0.56)	(0.05)	(0.47)	(0.59)	(1.43)	(0.62)

As shown above from the previous researches, different techniques have been used with different datasets (types of cancer). The above used classifiers have different performance result that may rely on preprocessing stages for datasets. Preprocessing stages represented in feature selection process may lead to improved performance for classifiers or lead to low performance. Datasets feature reduction is one of the main challenges that may face any classifier. Techniques for feature selection or reduction have shown that it can be a powerful assisted tool for classification. In addition the classification performance and percentage of accuracy may rely on the performance of the feature selection techniques. A comparative study is made to analyze research experiments. Table (9) shows the comparative study for the diagnosis and detection of cancer.

Table (9): A comparative study for the Stated Techniques above, for Classification and Clustering

S.no	Year	Dataset used	Techniques and algorithms used	Percentage of performance
1	2008	Colon cancer, Brain tumor, DLBCL	Artificial immune re Recognition System - Memory cell evolution and Somatic hyper mutation	Percentage of reduction in data sets used. The dimension reduction percentages are Colon: 93.30%, Brain Tumor: 73.20%, Nine Tumor: 97.15%, and DLBCL: 84.11%.
2	2012	Cancerous and normal Datasets. http://www.ncbi.nlm.nih.gov/geo/	Projection algorithm	Projection algorithm was compared with logistic regression, artificial neural networks, and Fisher discriminate function, it detect cancer correctly with percentage 89%, and other techniques respectively 82%, 78% and 75%
3	2010	Lymphoma dataset	SVM with Analysis of Variance (ANOVA)	SVM was sufficiently good classifier and In comparison with back propagation network, SVM gives result 97.91 and BPN 97.43
4	2010	Colon Tumor, (Central Nervous System) Tumor, (Diffuse Large B-Cell Lymphoma), Leukemia 1, AML, Lung Cancer, Prostate Cancer, Breast Cancer, and Leukemia 2	α Depended degree-based feature selection approach	Depend degree of feature is applied with different classifier successfully and Highest percentage has been shown
5	2012	Tumors samples from breast cells, blood cells, brain cells	fuzzy clustering with viewpoints	View point with the fuzzy cluster algorithm was compared with other three cluster algorithm (K-means, Fuzzy C-means, affinity propagation. Lower prediction error in different datasets than other techniques
6	2013	Colon dataset, CNS (Central Neural System) dataset, Lung cancer dataset as a binary class samples and Glioma dataset consists of four subtypes, as a multiclass dataset	Hybrid system between swarm intelligence algorithm ACO as a feature selection method and SVM as a classifier	By trying different number of features in many trials. It improves the SVM classifier in different datasets.
7	2013	Hepatitis Dataset, Liver Disorders Dataset. Diabetes Dataset.	Artificial bee colony (ABC) algorithm for feature selection and support vector machines for classification.	Proposed model yield 94.92%, 74.81%, and 79.29%, for the three datasets. Higher than other techniques
8	2013	Breast cancer	ACO as feature selection method based on rough set theory (RST)	Proposed model give the low number of informative genes
9	2013	Heart disease, lung dataset	Swarm intelligence algorithm called Bat, new algorithm called bat algorithm for attribute reduction BAAR	The proposed algorithm give the low number of informative attribute
10	2013	Diabetes microarray dataset	Hybrid system between ant colony optimization (ACO) and artificial bee colony (ABC) for feature selection, and technique called memory resistant inspired from fuzzy expert system as a classifier	Prove their effectiveness in generating rules that may in reducing features and improve help fuzzy systems in classification
11	2010	HIV, Cleveland heart, lung cancer, dermatology	Hybrid system between artificial bee colony (ABC) with rough set theory	The proposed model gives the lower number of informative features than other methods
12	2013	ALL, AML, COLON	Hybrid system between locality sensitive discriminant analysis(LSDA) and artificial bee colony (ABC) and fuzzy c means as a clustering	The proposed model gives the lower rate of error due to the low number of informative attributes
13	2013	Breast cancer, diabetes	Swarm intelligence technique called bat algorithm, modified algorithms called binary bat algorithm BBA, used as a feature selection and optimum path forest as a classifier	Low number of features is selected, which improve the classifier performance
14	2015	Colon, lukemial1, lung, SRBCT, lymphoma, leukemia2	Artificial bee colony (ABC) hybridized with method called minimum redundancy maximum relevance (mRMR) as a feature selection model, and SVM as a classifier	According to the performance of the feature selection model, SVM gives the best percentage in classification as shown in table (7)
15	2014	Brain tumor, leukemia, SRBCT, LUNG cancer, prostate tumor	Hybrid system between DT (decision tree) as a classifier and PSO (particle swarm optimization) as a gene selection method	Performance of PSODT prove that it can earn better percentage in classifying gene related to different kinds of cancer
16	2015	Breast cancer and its different status, lung dataset	Artificial Neural Network	Higher percentage 96% in ANN with three nodes and three hidden layers

5. CONCLUSION AND ANALYSIS

This survey shows a study on different existing techniques that have been applied on microarrays. Microarray as shown above has a big challenge called high dimensional data. This challenge has been solved using many techniques. As shown in this study high dimensional data problem is solved using feature selection methods. Many different existing gene selection methods have been used and proved their success in gene selection process. In which it will improve the classification process in order to classify cancerous or any other disease dataset with multi class or binary class. The problem is how to identify and detect different kinds of infected genes with different characteristics, such as mutated genes, or replicated genes. As known the mutations are caused by viruses, radiation, mutagenic chemicals. Most of the researchers depend mainly on how to employ machine learning techniques into the model has been built. In order to use machine learning techniques in analyzing microarray data, there must be kind of classifier with feature reduction method. The aim of the study is to show different hybridized methods that can eliminate noise, reduce data and classify this data. Swarm intelligence algorithms have been proved their performance, specifically in feature selection, which is considered as an important phase that may affect any classifier.

Swarm intelligence algorithms such as Ant Colony Optimization (ACO), Artificial Bee Colony optimization (ABC), Particle Swarm optimization (PSO) and other systems has been emerged to be applied on many fields such as biological field. Hybridization between popular machine learning techniques and emerged machine learning techniques such as swarm intelligence algorithms proved that it can yield us better results in diagnosis and classification. This study shows the related work of researchers and their contribution in using such intelligent systems like Artificial Bee Colony (ABC), ant colony optimization (ACO), particle swarm optimization (PSO) and other techniques, to hybridized one of these systems with known classifiers to get an improved result in classification. Researchers used swarm intelligence algorithms in feature selection and proved their capability in reducing features, to give the most informative features. Results showed that these techniques improve classification process.

Another aim of this study is to show that there is still a type of researches that is not done yet. Researchers hybridized computational methods with swarm intelligence (SI) methods, and prove that these hybridized systems give better results. Another trial must be done, is to build a model that relies on swarm intelligence algorithms. Clearly swarm intelligence can be used as a classifier as well as feature selection method. In another words the whole solution must rely on swarm intelligence algorithms.

6. REFERENCES

[1] Karmen Stankov," Bioinformatics tools for cancer geneticists", Arch Oncol, 2005; 13(2):69-75.

[2] P.K.Vaishali & Dr. A.vinayababu," Application of Microarray Technology and Soft computing in Cancer Biology: A Review", 2011, International Journal of Biometrics and Bioinformatics (IJBB), Volume (5): Issue (4): 225

[3] Martin Dufva,"Introduction to Microarray Technology", "DNA Microarrays for Biomedical Research: Methods and Protocols", vol. 529, 2009

[4] Yvan Saeys, I ˜naki Inza and Pedro Larra˜naga," A review of feature selection techniques in bioinformatics", August 24, 2007, Vol. 23 no. 19 2007, pages 2507–2517

[5] Adi L. Tarca, Roberto Romero, Sorin Draghici," Analysis of microarray experiments of gene expression profiling", American Journal of Obstetrics and Gynecology (2006) 195, 373–88

[6] Erick Suárez, Ana Burguete, Geoffrey McLachlan," review article Microarray Data Analysis for Differential Expression: a Tutorial", PRHSJ Vol. 28 No. 2, June, 2009

[7] Rainer Breitling," Biological microarray interpretation: the rules of engagement", 2006 Jul; 1759(7):319-27. Epub 2006 Jul 13.

[8] Liang Lan, Slobodan Vucetic," Improving accuracy of microarray classification by a simple multi-task feature selection filter",Int. J. Data Mining and Bioinformatics, Vol. 5, No. 2, 2011

[9] Dorina Kabakchieva," Predicting Student Performance by Using Data Mining Methods for Classification", 2013, CYBERNETICS AND INFORMATION TECHNOLOGIES Volume 13, No 1

[10] K Raja Sekhar, V Srinivasa Kalyan, B Phanindra Kumar," Training of Artificial Neural Networks in Data Mining", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-3, Issue-2, and July 2013

[11] V. Anuja Kumari, R.Chitra ," Classification Of Diabetes Disease Using Support Vector Machine", International Journal of Engineering Research and Applications (IJERA) ,Vol. 3, Issue 2, March -April 2013, pp.1797-1801

[12] Xin-She Yang and Xingshi He, "Swarm Intelligence and Evolutionary Computation: Overview and Analysis", 2015, Recent Advances in Swarm Intelligence and Evolutionary Computation, Studies in Computational Intelligence

[13] Ms. T.Hashni , Ms .T.Amudha," Relative Study of CGS with ACO and BCO Swarm Intelligence Techniques",2012, Int.J.Computer Technology & Applications, Vol 3 (5), 1775-1781

[14] Adis ALIHODZIC Milan TUBA, "Framework for Bat Algorithm Optimization Metaheuristic", Recent Researches in Medicine, Biology and Bioscience, Bioscience 4th International Conference on Bioscience and Bioinformatics (ICBB '13), Chania, Crete Island, Greece; 08/2013

[15] M. Dorigo, V. Maniezzo, et A. Colorni, Ant system: optimization by a colony of cooperating agents, IEEE Transactions on Systems, Man, and Cybernetics--Part B , volume 26, num. 1, pages 29-41, 1996.

[16] Poonam Sehrawat, Harish Rohil, "Taxonomy of Swarm Optimization", Volume 3, Issue 8, August 2013, International Journal of Advanced Research in Computer Science and Software Engineering

[17] Chuanliang Chen, ChuanXu and RongfangBie, "Artificial Immune Recognition System for DNA Microarray Data Analysis", Natural Computation, 2008, ICNC '08. Fourth International Conference (Volume: 6),

Page(s): 633 – 637

[18] Nazario D. Ramirez-Beltran, Joan M. Castro, Harry Rodriguez," A Projection Algorithm to Detect Cancer Using Microarray", IAES International Journal of Artificial Intelligence (IJ-AI), Vol. 1, No. 2, June 2012, pp. 91~102

[19] A. Bharathi, Dr.A.M.Natarajan, "Cancer Classification of Bioinformatics data using ANOVA", June, 2010, International Journal of Computer Theory and Engineering, Vol. 2, No. 3, 1793-8201

[20] Xiaosheng Wang and Osamu Gotoh," A Robust Gene Selection Method for Microarray-based Cancer Classification", 2010, Cancer Informatics, pages15–30

[21] Katerina N. Karayianni, George M. Spyrou and Konstantina S. Nikita," Clustering Microarray Data using Fuzzy Clustering with Viewpoints " ,Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), November 2012

[22] Sudip Mandal and Indrojit Banerjee," Cancer Classification Using Neural Network", International Journal of Emerging Engineering Research and Technology, Volume 3, Issue 7, July 2015, PP 172-178

[23] Hualong Yu, n, Jun Ni, Jing Zhao," ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data", 2013, Neurocomputing, pages: 309-318

[24] Mustafa Serter Uzer, Nihat Yilmaz, Onur Inan,"Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification", 2013.

[25] Hesham Arafat, Rasheed M.Elawady, Sherif Barakat and Nora M.Elrashidy," Using Rough Set and Ant Colony optimization In Feature Selection", February 2013,volume 2,Issue 1.

[26] Ahmed Majid Taha, Alicia Y.C. Tang," Bat Algorithm for Rough Set Attribute Reduction", Journal of Theoretical and Applied Information Technology, 2013, vol 51, No 1, ISSN: 1992-8645

[27] P. Ganesh Kumar, S. Arul Antran Vijay, D.Devaraj," A Hybrid Colony Fuzzy System for Analyzing Diabetes Microarray Data", IEEE, 2013, Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)

[28] N. Suguna and Dr. K. Thanushkodi,"A Novel Rough Set Reduct Algorithm for Medical Domain Based on Bee Colony Optimization", 2010, VOLUME 2, ISSUE 6

[29] K.Sathishkumar, Dr.V.Thiagarasu, M.Ramalingam," AN EFFICIENT ARTIFICIAL BEE COLONY AND FUZZY C MEANS BASED CLUSTERING GENE EXPRESSION DATA",2013,vol1,issue 5, International Journal of Innovative Research in Computer and Communication Engineering.

[30] Rodrigo Yuji Mizobe Nakamura, Luı´s Augusto Martins Pereira, Douglas Rodrigues, Kelton Augusto Pontara Costa, Joa˜o Paulo Papa and Xin-She Yang," Binary Bat Algorithm for Feature Selection",2013, Swarm Intelligence and Bio-Inspired Computation.

[31] Hala Alshamlan, Ghada Badr and Yousef Alohali," mRMR-ABC: A Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling", 2015, Volume 2015, Article ID 604910, 15 pages, BioMed Research International.

[32] Kun-Huang Chen, Kung-JengWang, Min-Lung Tsai, Kung-Min Wang, Angelia Melani Adrian, Wei-Chung Cheng, Tzu-Sen Yang, Nai-Chia Teng, Kuo-Pin Tan and Ku-Shang Chang, "Gene selection for cancer identification: a decision tree model empowered by particle swarm optimization algorithm",2014,Chen et al. BMC Bioinformatics 2014, 15:49