Intelligent Decision Support System for Breast Cancer Diagnosis by Gene Expression Profiles

Hanaa Salem[1], Gamal Attiya[2], Nawal El-Fishawy[2]

[1] Communications & Computer Department, Faculty of Engineering, Delta University, Egypt.
[2] Computer Science & Engineering Dept., Faculty of Electronic Engineering, Menoufia University, Egypt.
hana_future@gmail.com, gamal.attiya@yahoo.com, nelfishawy@hotmail.com

Abstract: Breast cancer transpires as one of the leading cause of deathly diseases among women worldwide. Nevertheless, there is evidence that early detection and treatment can increase the survival rate of breast cancer patients. This paper presents an Intelligent Decision Support System (IDSS) for breast cancer diagnosis by using gene expression profiles. The proposed system first extracts significant features from the input patterns by using Information Gain (IG) and then employs Deep Genetic Algorithm (DGA) for feature reduction as well as for breast cancer diagnosis. The proposed system is evaluated by considering a benchmark microarray dataset and compared with the most recent systems. The results show that the proposed IDSS outperforms other systems in terms of diagnosis time and accuracy. The proposed system produces 100% classification accuracy. In addition, the proposed system reduces the required memory space.
[Hanaa Salem, Gamal Attiya, Nawal El-Fishawy. **Intelligent Decision Support System for Breast Cancer Diagnosis by Gene Expression Profiles.** *Cancer Biology* 2016;6(1):68-79]. ISSN: 2150-1041 (print); ISSN: 2150-105X (online). http://www.cancerbio.net. 12. doi:10.7537/marscbj06011612.

Keywords: Decision Support System, Breast Cancer Diagnosis, Genetic Algorithm, Information Gain, Feature Selection

1. Introduction

Breast cancer is one of the common cancer-related diseases among women worldwide. Invasive breast cancer happens in nearly one out of eight (12 %) women during their lifetime. On January 2012, more than 2.9 million US women with breast cancer were alive [1]. Some of these women were cancer free, while others still has evidence of cancer and may have been go through treatment. In 2013, an estimated 232,340 new cases of invasive breast cancer are diagnosed among women [2]. In 2015, the American Cancer Society's estimates that about 231,840 new cases of invasive breast cancer will be diagnosed in women, about 60,290 new cases of carcinoma in situ (CIS) will be diagnosed (CIS is non-invasive and is the earliest form of breast cancer) and about 40,290 women will die from breast cancer [3]. Although a very intensive research has been carried out, there is still no concrete evidence of the root cause, preventive methods and the much-anticipated cure for cancer [4]. In reality, some of the cancerous tissues appear to be very aggressive. Therefore, early detection and treatment of cancer minimize the risk for the cancerous tissue to spread to other organ. If the cancerous cells are diagnosed at a localized stage, the chance of survival is extremely high. Therefore, the early detection of breast cancer is the key to increase survival rate. The traditional method for diagnosing the disease relies on human skills to identify the occurrence of convinced pattern from the database. However, this age-old method may subject to human error, inaccurate, time-consuming and labor intensive, and cause unnecessary burden to radiologists. Moreover, by the time of the detection completed, it may already be at a critical stage [5]. Recently, a number of Computer Aided Diagnosis (CAD) systems and machines learning techniques have been developed and functional in order to help doctors in the diagnosis decision process. In these systems, several approaches have been used to detect and classify breast cancer, e.g., a Meta-learning method based on Grammar Evolution (MGE) [6], decision trees C4.5 algorithm, ID3 algorithm and CART algorithm to categorize these infections and match the effectiveness, correction rate between them [7]. A hybrid approach for automated diagnosis in medical genetics is visual diagnostic decision support system services machine learning (ML) algorithms and digital image processing techniques [8]. Other approaches are Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) [9].Recent advances in microarray technology have opened new research direction for medical diagnosis. Microarray gene expression technology has opened the possibility of investigating the activity of thousands of genes simultaneously. Gene expression profiles show the relative abundance of messenger ribonucleic acid (mRNA) corresponding to the genes. Gene expressions refer to the production level of proteins specific for a gene. Thus, discriminant analysis of microarray data has great potential as a medical diagnostic tool since results represent the state of a cell at the molecular level. The goal of microarray

data classification is to build an efficient model that identifies the differentially expressed genes and may be used to predict class membership for any unknown samples [5]. The application of microarray data for cancer type classification has recently gained in popularity. Several techniques have been used to implement feature selection, e.g., data mining and genetic algorithm [10], hybrid information gain and genetic filter/wrapper algorithm [11], multiple PCA with sparsity [12], decision rules (feature gene pairs) mining algorithm[13], genetic algorithm [14], discrete wavelet [15], mutual information [16], regularized least squares, entropy-based techniques, instance-based techniques, random forests, least squares support vector machines, and various clustering techniques such as K-means clustering [17].DNA microarray technology is a great platform efficiently used for the analysis of gene expression in a wide variety of experimental researches. However, because of the large number of features ((in the request of thousands) and the little number of samples (basically not as much as a hundred) in this type of datasets, microarray data analysis face the "large p-small n" paradigm also known as the execrate of dimensionality. In spite of the very intensive research effort, challenges posed in microarray classification are the availability of only a limited number of samples in comparison to the high-dimensionality of the samples, and experimental variations in measured gene expression levels [5]. The minor number of cancer samples typically available to train the model compared with the number of genes features can reduce the performance of the classifier and present the risk of over-fitting. Cancer classification based on gene expression data contains a great number of features, which needs a relatively large training set to learn a classifier with a small error rate. Over-fitting a classification approach may be avoided by selecting a subset of genes features to learn a model. This paper presents an Intelligent Decision Support System (IDSS) for breast cancer diagnosis by gene expression profiles. The proposed IDSS combines the Information Gain (IG) to and two stages Genetic Algorithm (GA) called Deep Genetic Algorithm (DGA). The system uses the Information Gain (IG) to extract significant features from the input patterns. Where, an information gain value is first calculated for each gene (feature), the features are then arranged according to the IG and finally the features are selected based on a predefined threshold. In addition, the system uses the DGA for feature reduction and breast cancer diagnosis. The DGA uses GA to first extract higher-level features from the input vectors (feature reduction), after which, these features are given to the main GA to-do the actual prediction by dividing selected features into two classes and comparing the summation of gene expression value in each class. The rest of this paper is organized as follows. Section 2 presents the literature survey of related work while Section 3gives an overview of the information entropy and the information gain. Section 4 presents the proposed intelligent decision support system and describes the workflow of the proposed system. Section 5 illustrates the experimental results while Section 6 presents the conclusion and the future progress of the research work.

2. Related Work

Several techniques have been developed and tuned aiming to early detect breast cancer. In [18], a knowledge selection and classification of breast cancer disease using Adaptive Neuro-Fuzzy Inference System (ANFIS) is developed. The ANFIS is evaluated by using the Wisconsin Breast Cancer Diagnosis (WBCD) dataset established at University of California, Irvine (UCI). The results show that the performance is improved and the accuracy of classification is98.25%. Intelligent system that includes the artificial neural networks (ANN) based expert system for the automatic breast cancer diagnosis is becoming popular among researchers. In [19], a numerous intelligent techniques covering supervised and unsupervised Artificial Neural Network (ANN), and statistical and decision tree based, have been applied to classify dataset related to breast cancer health care obtained from the UCI repository site. The individual approaches are first tested and then collective together to form ensemble approach. The experimental results show that the accuracy obtained by applying the ensemble approach is better than that obtained by applying the individual approaches. However, Counter Propagation Network (CPN) is a good approach among all other individual models. The obtained accuracy by this approach is very near to that obtained in case of ensemble approach. In [20], a thermal camera for imaging the patients, significant parameters was resulting from the images for their rearward analysis with the assistance of a genetic algorithm. A fuzzy neural network was passed with the principal components for clustering breast cancer was identified. The number of images used for the test included a database of 200 patients out of whom15 were diagnosed with breast cancer via mammography. Results of the base method appearance a sensitivity of 93%.Training of the fuzzy-neural network was of the order of clustering $1.0923><10^{-5}$, reached to 2% and the selection of parameters in the hybrid module gave rise-measured errors. In [21], a medical decision support system is developed based on Genetic Algorithm (GA) and Least Square Support Vector Machine (LS-SVM) for the detection diabetes on a Pima Indian Diabetes

database of UCI machine learning repository. The system uses Genetic algorithm to select additional significant feature subset from the given feature set of the database and uses Least Square Support Vector Machine for classification. The performance of the suggested system is analyzed using various parameters like classification accuracy, using 10-fold cross-validation and distraction matrix. The results show that the classification accuracy of the suggested system outperforms that of different standing systems. The accuracy of the system for the PID database was found to be 81.33% with GA as a feature selection method. In [22], classification of cancer based on gene expression has provided insight into possible treatment strategies. Supervised learning techniques that have been active to classify cancers, a hybrid feature selection method based on an attribute selection method Relie IF and a genetic algorithm used to find a set of genes that can best distinguish between cancer subtypes or normal against cancer samples. The application of various classification methods (decision tree, k-nearest neighbor, support vector machine, bagging, and random forest) on 5 cancer databases shows that no classification process generally outmatch all the others. However, the k-nearest neighbor and linear SVM improve the classification performance over other classifiers. In [23], Gene range select based on a random forest method lets selective subset for better classification of cancer databases was proposed. Results show's that various gene arrays assist in increasing the overall classification accuracy of the cancer related databases, as the amount of genes can be further scrutinized to form the best subset of genes. It can support the gene-filtering technique for further analysis of the microarray data in gene network analysis, gene-gene interaction analysis and several other related fields. However, as the number of genes features and information rise, it becomes more interesting to integrate the disparate database into a reliable classification model. Enhancing machine-learning techniques that can successfully distinguish among cancer subtypes or normal versus cancer samples is vital.

3. Entropy and Information Gain

Naturally, gene expression dataset keep a high dimension and a small sample size. This makes testing and training of general classification methods very hard. In general, only a relatively small number of gene expression data out of the total number of genes considered shows a significant correlation with a certain phenotype. In other words, even though thousands of genes are usually investigated, only very small number of these genes displays a correlation with the phenotype in question. Therefore, in order to study gene expression profiles correctly, feature selection (also called gene selection) is crucial for the classification process [24]. Feature selection depends on the importance of the numerous features, characteristics after removing redundant distinct features, picking out the classification of certain significant features to reduce the dimension of the feature space [25].Entropy is a basic vital concept in information theory. Shannon [26] uses the concept of entropy in information processing and offered the concept of 'information entropy'. Entropy is a measure for computing information and is a measure of the degree of uncertainty of a random variable. In the information gain, the measure of the importance of the feature is to see how much information could classify as to bring the more information, the more important features[27]. If X is a discrete random variable with probability function, its entropy is defined by:

$$H(X)= - E\mathbf{i}P(X\mathbf{i}) 1 o g_{2}(P(_{Xi})) \quad (1)$$

It is seen that, extra changes of random variables, greater information obtained through them. For the classification system, class C is variable, so the entropy of the classification system can be defined as:

$$H(C)= - £^{L}_{= x}P(C\mathbf{i}) 1 og\mathbf{2}(P(q)) \quad (2)$$

Here, P(ci) represents priori probability of the categorical variables C and is the categories number of the classification system. In particular, for two classification problems (where L number of classes, L=2), information entropy in equation (2) can be defined as:

$$H(C) = - \mathbf{P(cJ} 1og\mathbf{2}(P(c_{x})) - Pfe) 1og\mathbf{2}(Pfe)) \quad (3)$$

For a gene X, it may have n possible values (xj, x_2, x_n). The corresponding conditional entropy is

$$\mathbf{H(^{c}/x)} = - Z?= \mathbf{i} P(X) £^{L}_{=} \mathbf{i} P(C / X) 1og\mathbf{2} (P(C/ \mathbf{x})) \quad (4)$$

$P(c_i/x_J)$ represents the conditional probability of variables C after gene X is fixed, n all number of genes and L number of classes. Thus, the Information Gain (IG) of gene X brings to the classification system can be expressed as the difference between the original system information entropy and the conditional entropy after gene X is fixed.

$$IG(X) = H(C) - H(C/X) \quad (5)$$

If gene X and category C are not relevant IG(X) = H(C) - H(C/X) = zero. While, if relevant, H(C) > H(C/X), i.e., IG(X) = H(C) - H(C/X) > 0. The larger the difference is, the stronger the correlation between X and C. Therefore, the differential entropy defined as information gain, represents the amount of information obtained after the elimination of uncertainty. Clearly, larger information gain value a feature item has, the larger involvement it makes, extra important for the classification. Therefore, when choosing genes, usually choose genes with great information gain to signify the original

high-dimensional gene first, and use them as an origin for further gene selection.

4. Proposed System

Figure 1 shows the general framework of the proposed Intelligent Decision Support System (IDSS). The system first accepts Gene Microarray Dataset as input patterns, then selects significant features (feature selection) from the input patterns by using Information Gain (IG) and finally the system employs two stages genetic algorithm, called Deep Genetic Algorithm (DGA), for data reduction as well as for breast cancer diagnosis.

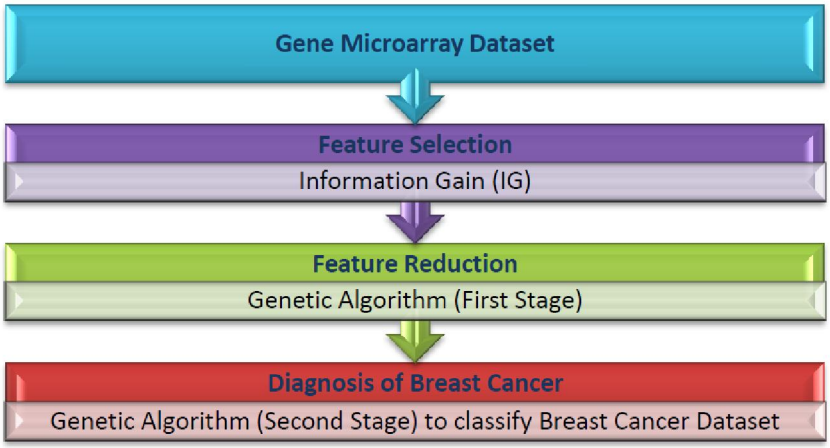

Figure 1: Proposed IDSS Framework

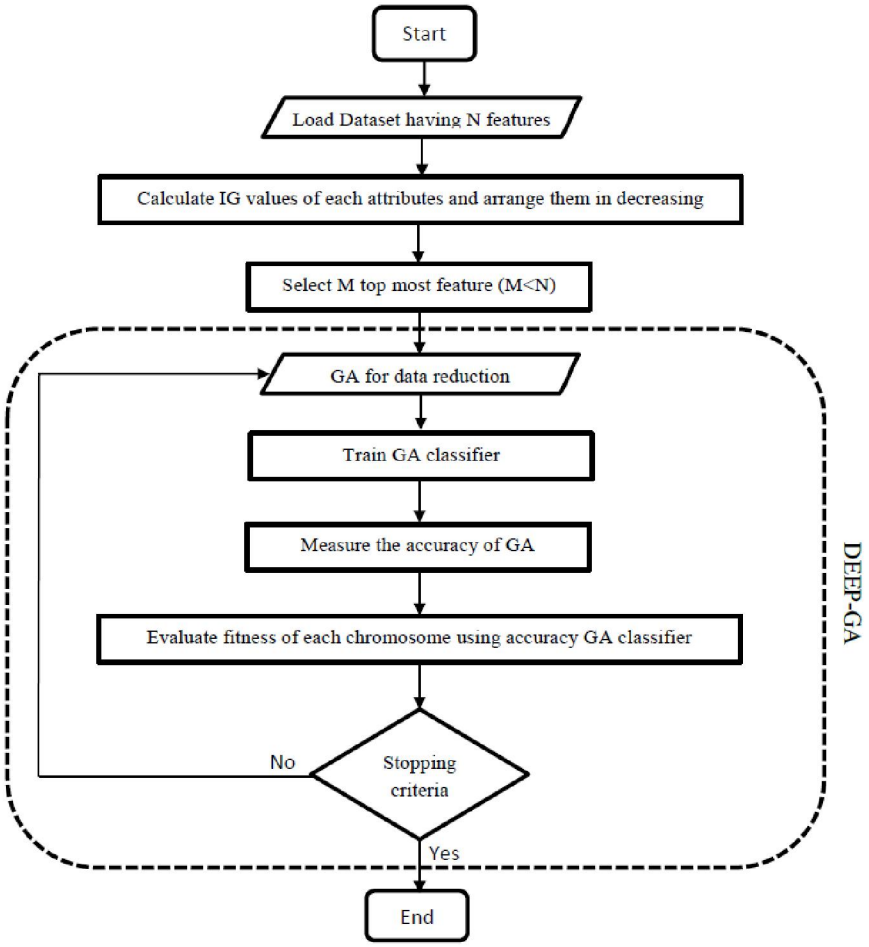

Figure 2: Workflow of the proposed system

4.1 Proposed System Workflow

Figure 2 shows the workflow of the proposed system. The system works as follows:

1. Load dataset having N attributes.
2. Calculate Information Gain (IG) value of each gene.
3. Arrange attributes (gene features) in decreasing order according to their IG values.
4. Select the M top most attributes (M<N) whose IG value is greater than a predefined threshold value.
5. Initiate parameters of Genetic Algorithm (GA) like population size, crossover rate, and mutation rate.
6. Create population of attribute.
7. Train the classifier by the resulting chromosomes (feature subset).
8. Measure the accuracy of GA classifier.
9. Find the fitness value of each chromosome using accuracy function of genetic classifier.
10. Apply crossover and mutation for generation of new chromosomes while stopping criterion is not valid.
11. Repeat step 7 and 9 while stopping criteria do not meet.

4.2 Gene Microarray Dataset

The microarrays datasets are managed as a matrix with its rows denote the genes (features) and the columns denote the samples. Commonly, only a small number of gene expression data display a strong correlation with a certain phenotype matched to the whole number of genes investigated. This implies that if a large number of genes studied; only a small number show significant correlation with a certain phenotype. Therefore, in order to analyze gene expression profiles validly, feature (gene) selection is crucial for the classification process. The objective of feature selection is to recognize the subset of differentially expressed genes that are potentially appropriate for distinguishing the sample classes [28].The training data matrix will be utilized for the gene selection step and the result diminished train subset will train the implemented classifiers. The test data matrix will be the controller to see if the suggested classification systems are effective by analyzing what number of test samples they will classify accurate [29].

4.3 Information Gain Algorithm

The information gain algorithm works as follows: Input: original gene sets C; Output: selected gene subset feature selection (FS).

1) Establish Classification Attribute (in Table 1).

2) For each class of known samples probability, compute classification entropy according to the probability

3) using the formula (2)

4) For each attribute (gene) in table 1, calculate the probability of all of its values. Compute conditional

5) Probabilities.

6) According to the probability obtained using the formula (4) for each gene (attribute), calculates conditional

7) Entropy.

8) Calculate Information Gain using classification attribute (5) for all genes (attributes).

9) Sort the outcomes obtained in 5) and Select k Attribute with the highest gain as a compact subset of genes FS

10) (Depends on threshold).

4.4 Deep Genetic Algorithm

The Deep Genetic Algorithm (DGA) consists of two stages of genetic algorithm. At the first stage, the genetic algorithm learns to extract relevant features from the input patterns or from the extracted features by the IG. At the first stage, GA performs the actual prediction of breast cancer diagnosis using significant extracted features as inputs.

4.4.1 Genetic Algorithm based Feature Reduction (First Stage)

In this stage, the features chosen by IG are used for feature selection by the genetic algorithm GA. Figure 3 shows the GA procedure. The population is initialized randomly, with each chromosome in the population coded to a binary string. The chromosome length represents the number of the features. The bit value {1} represents a selected feature, whereas the bit value {0} represents a non-selected feature. Standard genetic operators, such as crossover and mutation, are applied without modification [30].

a. Encoding and Initial Population

Each variable is connected with one bit in the string. If the i^{th} bit is active (value 1), then the i* gene is selected in the chromosome. While, a value 0 indicates that the corresponding feature is ignored. In this way, each chromosome represents a dissimilar feature subset.

b. Selection

Roulette wheel selection is used to probabilistically choose the individual to practice a parent mating pool which size is alike to the population size minus the elitism number. The probability that an i^{th} individual is selected is given by(6) ***Li*** ^{F_i}iHere F_i and PopSize are the fitness of i^{th} individual and the population size respectively. In this way, the fitter individual will have a good chance to be selected for intermarriage and thus will inherit their genetic information in the next generation.

c. Crossover

The first and second individuals from the intermarriage pool are paired for the crossover operation. This is trailed byte third and fourth

chromosomes and the process is continual until the last and second last chromosomes. If the size of the parent pool is odd, the first chromosome is moved to the temporary population before pairing the remaining. Crossover is used to swap the genetic material of chromosomes between selected couples to produce new offspring that are capable of preserving the characteristics of the parent chromosomes well. Many kinds of crossover procedures have been tried in GAs to date. In this study, a 2-point crossover operator was used, which chose two cutting points at random and alternately copied single segments out of each parent. The crossover rate was 0.8.

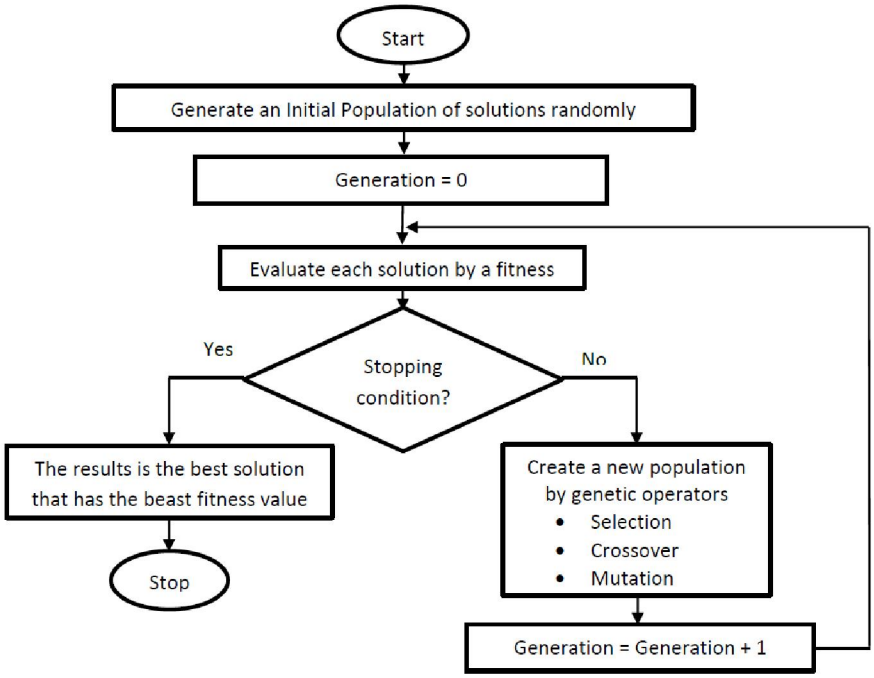

Figure 3: Genetic Algorithm

d. Mutation

After that, all the chromosomes resulted from the crossover will go through a mutation operation and consequently, anew offspring is produced. If a mutation was present, either one of the offspring's was mutated, and its binary representation changed from 1 to 0, or from 0 to 1 after the crossover operator is applied. If the mutated chromosome was superior to both parents, it replaced the worst chromosome of the parents; otherwise, the most inferior chromosome in the population was replaced [31]. The GA was configured to contain 100 populations and was run for20 generations in each configuration. The mutation rate was 0.1.

e. Fitness function

The fitness function assesses each chromosome in the population so that it may be ranked against all the other chromosomes. The essential goal of feature subset selection is to use little features to achieve the same or enhanced performance. Additionally, it has been found that the collection of features with low redundancy among them, by given that various information about the target class, and with a certain resemblance to the target class, can improve the performance rates. Hence, the fitness function should contain three terms: the misclassification error, the number of features selected and a redundancy measure among them. The purpose of the genetic search in the DGA approach is to seek "good" gene subsets having the minimal size and the highest prediction accuracy. To achieve this objective, we devise a fitness function taking into account this criterion.

f. Features Selection Procedures

Step 1: Generate random population of Y Chromosomes. These Chromosomes are potential solution for the given problem.

Step 2: Assess the fitness function f(x) of each chromosome x in the random population.

Step 3: Create a new population of Chromosomes by repeating the following steps until the new population is complete

3. a Select two parent chromosomes from the current population according to their fitness value (the best fitness, the larger chance to be selected)

3. b With a crossover probability cross over the parents to form a new offspring (children). If crossover operation was not complete, offspring is an exact copy of parents.

3. c With a mutation probability mutate new offspring at each locus (position in chromosome).
3. d Place new offspring in a new population
Step 4: Use newly generated population for the further runs
Step 5: If the end condition is satisfied, stop the process and return the best solution in current population
Step 6: Go to step 2.

4.4.2 GA-Based Classifier (Second Stage)

4.4.2.1 Genetic Programming (GP)

In fact, GP is a branch of genetic algorithm (GA), and the main difference between GP and GA is the structure of individuals. GA has string-structured individuals while GP's individuals are trees [32]. The GP structure is as follows:

- **Generate an introductory population of solutions:** The initial solutions are made to satisfy the population. There will be an expansive variety of solution structures through the procedure of this arbitrary generation. Figure 4 shows the solutions in the population of GP.
- **Evaluate every solution by a fitness function:** every solution is assessed to decide its fitness. The evaluation function, called "fitness function", is an imperative component GP. The fitness function is problem specific. Each solution will have a measure of goodness attendant with it.
- **Create a novel population by genetic operators:** The target of applying genetic operations on the population is to build the better quality population of the solutions. There are three genetic operators: reproduction, crossover, and mutation [33].

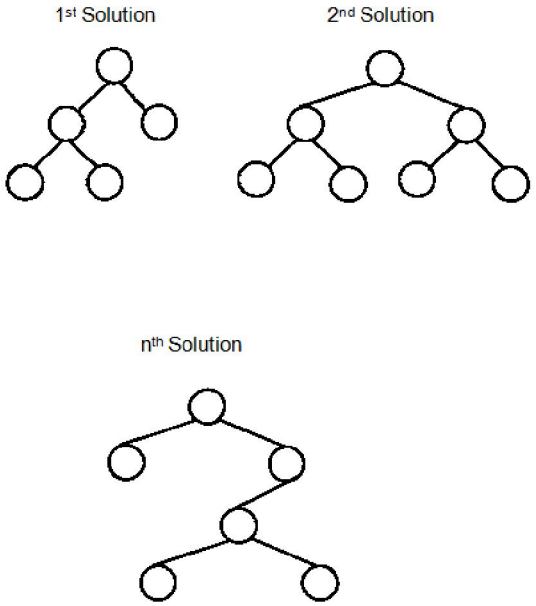

Figure 4: The Solution Structure of GP.

- **Reproduction:** An amount of good solutions are selected in view of their fitness value to be replicated to the next generation. This procedure saves good solutions.
- **Crossover:** This operator recombines parts from two great solutions, called "parents", to make new solutions, called "offspring" or "child". Two great solutions are chosen. The likelihood of a solution being chosen is corresponding to its fitness. The crossover points, which decide the location to exchange parts, are randomly selected. In GP, the sub-trees from parents are exchanged as shown in Figure 5. This process creates two new offspring [34].

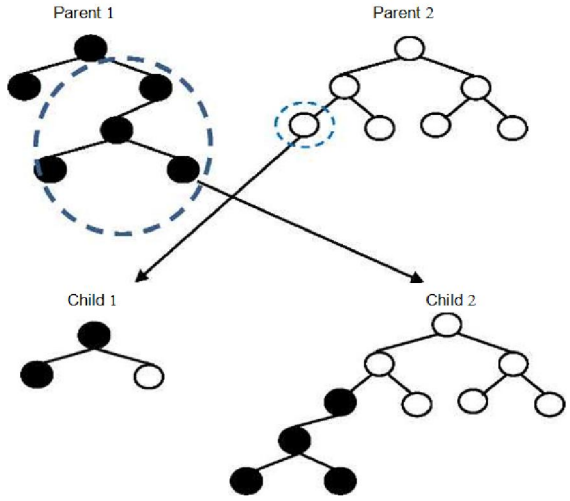

Figure 5. The crossover operator in GP

- **Mutation:** To keep diversity in the population and to empower investigation of distinctive solutions, the mutation operator changes some piece of a solution arbitrarily. A solution is selected randomly and a location to be changed is selected [35]. In GP, a part is mutated by supplanting it with a small arbitrary tree as shown in Figure 6.

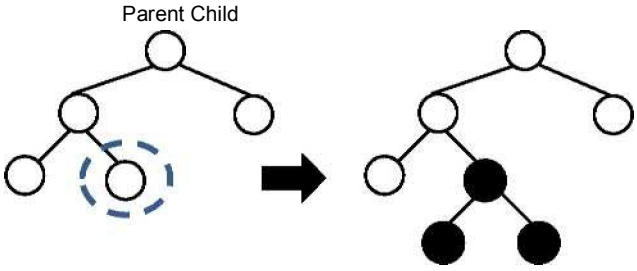

Figure 6. The mutation operator in GP

These stages have been repeated until the termination criteria are met. The end measure for the run may be characterized by the best fitness quality or

a most extreme number of generations. All through generations, the quality of solutions is moved forward and improved. The outcomes from every run are distinctive as the search for a solution is probabilistic and the solution for this problem is not one of a kind.**4.4.2.2 Classification by Means of GP**

GP-based classifier is denoted by a classification tree. It comprises of symbols from the function set F and the terminal set T. The function set F comprises of arithmetic operators and the terminal set T comprises of number of genes constants and variables characterized as takes after: F = {+, *, /} and T = {0... number of genes, xi...x_n}. The variables denote the value of the expression level of genes. To calculate the fitness of a candidate, its expression is assessed. The variables (xi... x_n) are data from the microarray dataset. If the result of evaluating an expression is more than 0, it is classified as Class 1. Otherwise it's classified as Class 2. An expression is assessed with data in the training set. The total number of the accurate classification is counted to calculate the fitness value of the expression. The higher fitness value indicates the better solution.

5. Experimental Results and Discussion

The proposed system is evaluated by using the Skewed cancer gene expression datasets downloaded from the Kent Ridge Bio-medical Dataset website [36].

5.1 Microarray Datasets

The Kent Ridge Bio-medical Dataset is an online repository of high-dimensional biomedical datasets including gene expression data, protein-profiling data and genomic sequence data that are related to classification and that are published recently in Science, Nature and so on prestigious journals. The database contains prognosis results of 78patients out of which 34 are relapse and 44 are non-relapse for train [36]. The microarrays dataset is arranged as a matrix. The rows of the matrix represent the genes (features) while the columns represent the samples (patients). The microarray gene expression data matrix Xij takes the size (mg * ms). Where, mg is the total number of genes ranges from 0 to i and m is the total number of samples ranges from 0 to j. As the number of collected samples are limited, the microarray data matrix partitioned into two matrices; training data matrix and testing data matrix. The training data matrix will be used for the gene selection stage and the result decreased train subset will train the implemented classifiers. The testing data matrix will be the guide of evaluating the proposed classification system, by noting the number of test samples that the system will classify correctly. Table 1 summarizes detailed information about the microarray datasets. The datasets have two classes; relapse and non-relapse, 78 training samples, 19 testing samples, 24481 genes, and 0.75-1.7 imbalance ratios.

Table 1: Dataset Details

Datasets	**Classes**	**Genes**	**Train Samples**	**Test samples**
Breast	relapse,	24481	78 (34 relapse & 44	19 (12 relapse & 7
Cancer	non-relapse (2)		non-relapse)	non-relapse)

5.2 Performance Metrics

Table 2 summarizes the various performance metrics. The results are measured in contradiction of the following diagnostic performance measures. True Positive (TP): the number of positive cases correctly detected. True Negative (TN): the number of negative cases correctly detected. False Positive (FP): the number of negative cases diagnosed as positive. False Negative (FN): the number of positive cases diagnosed as negative.

Table 2: Diagnostic performance measures Breast Cancer

Cancer Test	**Present**	**Absent**	**Total**
Positive	True Positive (TP)	False Positive (FP)	(TP+FP)
Negative	False Negative (FN)	True Negative (TN)	(TN+FN)
Total	(TP + FN)	(TN+FP)	(TP+FN+TN+FP)

These performance metrics are first computed and then used to compute Classification Accuracy (CA) of the algorithm according to equation (7).

$$CA = \frac{\text{No. of correct classified sampels (TP+TN)}}{\text{Total no. of samples (TP+FN+TN+FP)}} \quad (7)$$

5.3 Threshold Value

In the proposed system, the first step uses IG for feature selection. Each feature has its own IG value which regulates whether this feature is to be selected or not. The threshold value is applied for checking the features. If a feature has IG value greater than the predefined threshold, the feature is selected; otherwise, it is not selected. Greater information gain will result in a higher likelihood of gaining pure classes in a target class. After calculating the information gain values for all features, a threshold for the results was recognized.

Since most papers show that most IG values are zero after the computation process, not many features have an effect on the category in a data set, signifying that these features are irrelevant for classification. The thresholds in studies were 0 for most of the data sets. Table 3 shows that attribute ranking with threshold for discarding attributes: threshold value = 0.

5.4 GA Parameter Settings

The GA, a wrapper method is implemented. The features selected during the main-stage were used for feature selection by the genetic algorithm. The GA population is set randomly, with each chromosome in the population coded to a binary string. The bit value {1} signifies a selected feature, whereas the bit value {0} signifies a non-selected feature, however the chromosome length represents the number of the features. Both, the active features and the number of them are generated randomly. In experimentation, we use the population size of 100 individuals. Scattered crossover, in which each bit of the offspring is chosen randomly, was the choice for combining parents of the previous generation. The crossover rate was set to 0.8. Strategy selection based on roulette wheel and uniform sampling was Applied, excellent an elite count value of 50(number of chromosomes which are taken in the next generation). According to that, consider a traditional mutation operator which flips a specific bit with a probability rate of 0.2. A modification which includes mutating a random number of bits between 1 and the number of active features of the individual is presented. Since it was empirically proved that the best subsets include few features, this change avoids the increase on the number of most active attributes (features) in the last generations of the GA. Fitness functions will be used; Classification accuracy (CA) of genetic algorithm according to equation (7) and the used number of genes.

Table 3: Information Gain Ranking Filter; Attribute ranking with threshold for discarding attributes: threshold value = 0.

Datasets	No. of Instances	No. of Features	Gene Name and Gene No of Highest and Lowest Ranked Features		IG value	No. of features that has IG value greater than threshold 0
Breast Cancer-Training	78	24481	Highest Ranked	Contig7258 RC (377)	0.374	716
			Lowest Ranked	U58033 (4569)	0.116	
Breast Cancer-Testing	19	24481	Highest Ranked	NM 014835 (14724)	0.949	828
			Lowest Ranked	Contig22379 RC (6681)	0.38	

5.5 Experimental Results

Table 4 shows the experimental results of applying the proposed system on the breast cancer microarray gene expression dataset. From the table, IG threshold value 0.7 is the optimal value for this dataset. At this value, features are reduced from 24481 attributes to 45 attributes in IG and reduced farther to 22 features by applying GA with 100 population size and 20 evaluation progress. In addition, the accuracy of classification is 100%. By using the proposed system, memory space occupied by irrelevant and redundant attributes are removed and hence lot of memory space is reduced.GP method results in balanced and unbalanced trees with several different depths. The max depth of each individual is restricted to 3, and each non terminal is forced to have exactly three children, which can be terminals or non- terminals. The crossover rate is 0.8, and the mutation rate is 0.4. The population size is 100,and the maximum number of generations is 20. The fitness function for each individual is its accuracy on the validation set. The implementation of GP is based on the Pyevolve library [37].

Table 4: Classification accuracy and extracted features under different IG threshold values

IG Threshold Value	No. of features (Genes) After IG	No. of features (Genes) After GA	Accuracy of Classification
0.0	**828**	**396**	**78.9474 %**
0.38	**605**	**316**	**89.4737%**
0.5	**203**	**115**	**89.4737%**
0.7	**45**	**22**	**100%**
0.9	**18**	**8**	**94.7368 %**

In the current study, many thresholds are tested. For each threshold, if the information gain value of the feature was higher than the predefined threshold, the feature is selected; if not, the feature was not selected. From this study, the best threshold value for this dataset is 0.7, where it achieves accuracy of classification of 100%.

Figure 7 shows the classification accuracy of the proposed system (IG-DGA) and six different algorithms(GAANNRP, GAANNLM, GAANNGD and GA-LDA, GA-SVM and GA-NB) reported in [4, 38]. In the GAANNRP, GAANN LM, and GAANN GD algorithms [4], the feature selection method is Genetic Algorithm (GA) while the classifiers are Artificial Neural Networks (ANN) with three different variations of Back Propagation (BP) techniques. The different Back Propagation (BP) variations are namely Resilient Back-Propagation (RP), Levenberg Marquart (LM) and Gradient Descent (GD) with momentum. These BP variations are used for parameter optimization of the artificial neural networks (ANN) by fine-tuning of the weight of the ANN. In the GA-LDA, GA-SVM and GA-BN [38], the feature selection method is Genetic Algorithm (GA) while the classifiers are Linear Discriminant Analysis (LDA), Support Vector Machines (SVM) and Naive Bayes (NB). From Figure 7, by comparing the experimental results, the proposed system improves the sample classification accuracy; the accuracy of classification is 100%. The experimental results show that the proposed strategy is able to improve the stability of the selection results as well as the sample classification accuracy.

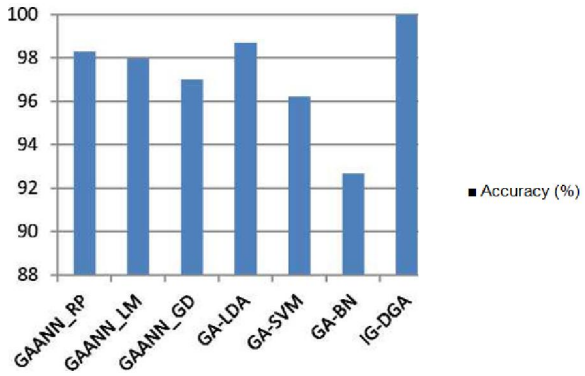

Figure 7: Classification Accuracy of the Proposed Method and Six Other methods.

6. Conclusions and Future Works

In this paper, an intelligent decision support system (IDSS) is developed based on Information Gain (IG) and Deep Genetic Algorithm (DGA). In the proposed system, the IG is employed to pre-select features while the DGA is used to further identify a small feature subset for accurate sample classification. A benchmark microarray dataset is used to evaluate the proposed algorithm. The experimental results suggest that the proposed strategy is able to improve the stability of the selection results as well as the sample classification accuracy. The proposed system can achieve a good result in terms of classification accuracy comparing with other methods, decreasing medical errors, and minimizing life-threatening events caused by delayed or uninformed medical decisions. This algorithm achieves the optimal value for accuracy of classification percentage 100% from 24481attributes which are reduced to 45 attributes in IG, which used 100 population size and 10 evaluation progress for GA feature Selection attributes reduced to 22. Memory space occupied by irrelevant and redundant attributes are removed and hence lot of memory space is reduced. In the future work, we will integrate various kinds of genomic data (e.g., interaction of protein-protein dataset and gene expression profile) to increase and enhance the prediction accuracy as compared to using gene expression alone.

References

1. R. Siegel, C. DeSantis, K. Virgo, et al., "Cancer Treatment and Survivorship Statistics", CA: A Cancer Journal for Clinicians, Vol. 62, No. 4, pp. 220-41, Jul-Aug 2012.
2. C. DeSantis, R. Siegel and A. Jemal, "Breast Cancer Facts & Figures 2013-2014", The American Cancer Society, Atlanta, Georgia, pp. 1-30, 2014.
3. http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics Accessed 20 July 2015.
4. F. Ahmad, N. A. M. Isa, M. H. M. Noor, and Z. Hussain, "Intelligent Breast Cancer Diagnosis Using HybridGA-ANN", Proceedings of the 5th International Conference on Computational Intelligence, Communication Systems and Networks (CICSyN), IEEE Computer Society, pp. 9-12, 2013.
5. F. Ahmad, N. A. M. Isa, Z. Hussain, M. K. Osman, and S. N. Sulaiman, "A GA-based Feature Selection and Parameter Optimization of an ANN in Diagnosing Breast Cancer", Pattern Analysis and Applications, Vol. 17,No. 2, May 2014.
6. F. Herrera, E. G. Rib'e and E. B. Mansilla, "Decision Support System for the Breast Cancer Diagnosis by AMeta-learning Approach Based on Grammar Evolution", The 9th International Conference on Enterprise Information Systems,

Decision Support Systems - Data Mining & Machine learning, Health - DSS - Decision Support Systems, pp. 222-227, 2006.
7. D.S. Kumar, G. Sathyadevi and S. Sivanesh, "Decision Support System for Medical Diagnosis Using Data Mining", International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, pp. 147-153, May 2011.
8. K. Kuru, M. Niranjan, Y. Tunca, E. Osvank, and T. Azim, "Biomedical Visual Data Analysis to Build an Intelligent Diagnostic Decision Support System in Medical Genetics", Artificial Intelligence in Medicine, Elsevier, Vol. 62, pp. 105-118, 2014.
9. K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine Learning Applications in Cancer Prognosis and Prediction", Computational and Structural Biotechnology Journal, Elsevier, Vol. 13, pp. 8-17, 2015.
10. Sand A. Kusiak," Data mining and genetic algorithm based gene/SNP selection", Artificial Intelligence in Medicine, Vol. 31, Issue 3, pp. 183-196, 2004.
11. C. H. Yang, L. Y. Chuang and C. H. Yang, "IG-GA: A Hybrid Filter/Wrapper Method for Feature Selection of Microarray Data", Journal of Medical and Biological Engineering, Vol. 30, pp. 23-28, 2010.
12. Y. Huang and L. Zhang, "Gene Selection for Classifications Using Multiple PCA with Sparsity", Tsinghua Science and Technology, Vol. 17, No. 6, pp. 659-665, December 2012.
13. H. Yu, Jun Ni, Y. Dan and S. Xu, "Mining and Integrating Reliable Decision Rules for Imbalanced Cancer Gene Expression Data Sets", Tsinghua Science and Technology, Vol. 17, No. 6, pp. 666-673, December 2012.
14. G. Chakraborty and B. Chakraborty, "Multi-objective Optimization Using Pareto GA for Gene-Selection from Microarray Data for Disease Classification", Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2629-2634, 2013.
15. J. Bennet, C. A. Ganaprakasam and K. Arputharaj, "A Discrete Wavelet Based Feature Extraction and Hybrid Classification Technique for Microarray Data Analysis", Hindawi Publishing Corporation, Scientific World Journal, pp. 1-9, 2014.
16. N. Hoquea, D. K. Bhattacharyyaa and J. K. Kalitab, "MIFS-ND: A Mutual Information-based Feature Selection Method", Expert systems with applications, Elsevier, pp. 1 - 25, 2014.
17. E. Naghieh and Y. Peng, "Microarray Gene Expression Data Mining: Clustering Analysis Review", IJCTT, Vol.3, pp. 387-390, 2012.
18. Fatima and C. M. Amine, "A Neuro-Fuzzy Inference Model for Breast Cancer Recognition", International Journal of Computer Science & Information Technology (IJCSIT), Vol. 4, No 5, pp. 163-173, October 2012.
19. H.S. Hota, "Diagnosis of Breast Cancer Using Intelligent Techniques", International Journal of Emerging Science and Engineering, Vol. 1, Issue 3, pp. 45-53, January 2013.
20. H. G. Zadeh, O. Pakdelazar, J. Haddadnia, G. R. Rad, and M. M. Zadeh, "Diagnosing Breast Cancer with the Aidof Fuzzy Logic Based on Data Mining of a Genetic Algorithm in Infrared Images", Middle East Journal of Cancer, Vol. 3, pp. 119-129, 2011.
21. S. Aishwarya and S. Anto, "A Medical Decision Support System based on Genetic Algorithm and Least Square Support Vector Machine for Diabetes Disease Diagnosis", International Journal of Engineering Sciences &Research Technology (IJESRT), Vol. 3, pp. 4042-4046, April 2014.
22. H. Hijazi and C. Chan, "A Classification Framework Applied to Cancer Gene Expression Profiles", J Healthc Eng, NCBI, Vol. 4, pp. 255-283, 2013.
23. K. Moorthy, M. S. B. Mohamad, and S. Deris, "Intelligent Information and Database Systems, Lecture Notes in Computer Science, Springer, Vol. 7802, pp. 385-393, 2013.
24. L. Y. Chuang, C. Hsuan Ke, and C. H. Yang, "A Hybrid Both Filter and Wrapper Feature Selection Method for Microarray Classification", Proceedings of the International Multi Conference of Engineers and Computer Scientists, Vol. 1, pp. 146-150, 2008.
25. W. Sha-Sha, L. Hui-Juan, J. Wei and L. Chao, "A Construction Method of Gene Expression Data Based on Information Gain and Extreme Learning Machine Classifier on Cloud Platform", International Journal of Database Theory and Application, Vol. 7, No.2, pp. 99-108, 2014.
26. J. C. Baez, T. Fritz and T. Leinster "A Characterization of Entropy in Terms of Information Loss", Entropy, Vol. 13, No. 11, pp. 1945-1957, 2011.
27. L. Chen, K. Wu and Y. Li, "A Load Balancing Algorithm Based on Maximum Entropy Methods in Homogeneous Clusters", International and Interdisciplinary open access Journal of Entropy and Information Studies, Vol. 16, pp. 5677-5697, 2014.

28. G. V. Sudha George and V. C. Raj," Review on Feature Selection Techniques and the Impact of SVM for Cancer Classification using Gene Expression Profile", International Journal of Computer Science & Engineering Survey (IJCSES). Vol.2, No.3, pp. 26-38, August 2011.
29. P. K. Ammu, V. Preeja," Review on Feature Selection Techniques of DNA Microarray Data", International Journal of Computer Applications, Vol. 61, No. 12, pp. 39-44, January 2013.
30. S. Shah and A. Kusiak," Cancer gene search with data-mining and genetic algorithms', Computers in Biology and Medicine, Elsevier, Vol. 37, pp. 251 - 261, 2007.
31. D. A. Salem, R. A. Abul Seoud, and H. A. Ali, "K5 Merging Genetic Algorithm with Different Classifiers for Cancer Classification using Microarrays", Proceedings of the 29th National Radio Science Conference, pp. 659 -666, 2012.
32. K. Oyebode and J. A. Adeyemo," Genetic Programming: Principles, Applications and Opportunities for Hydrological Modelling", World Academy of Science, Engineering and Technology International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering, Vol. 8, No. 6, pp. 348-354, 2014.
33. J. Eggermont, J. N. Kok and W. A. Kosters," Genetic Programming for Data Classification: Partitioning the Search Space", http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.9.8725.
34. V. B. Canedo, N. S. Marono, and A. A. Betanzos, "An ensemble of filters and classifiers for microarraydata classification," Pattern Recognition, Vol. 45, No. 1, pp. 531-539, 2012.
35. K. H. Liu, M. Tong, S. T. Xie and V. T. Y. Ng," Genetic Programming Based Ensemble System for Microarray Data Classification", Hindawi Publishing Corporation, Computational and Mathematical Methods in Medicine, Volume 2015, pp. 1-11, 2015.
36. http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html. Accessed 26 May 2015.
37. C. S. Perone, "Pyevolve: a python open-source framework for genetic algorithms," ACM SIGEVOlution, Vol. 4, no. 1, pp. 12-20, 2009.
38. R. M. Luque-Baena, D. Urda, J.L. Subirats, L. Franco, and J.M. Jerez," Analysis of Cancer Microarray Data using Constructive Neural Networks and Genetic Algorithms", Theoretical Biology and Medical Modelling, Vol. 11, pp. 1-18, 2014.

3/16/2016