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

Hanaa Salem[1], Gamal Attiya[2], Nawal El-Fishawy[2]
[1]Faculty of Engineering, Delta University, Gamasa, Egypt
[2]Faculty of Electronic Engineering, Menoufia University, Monouf, Egypt

ABSTRACT

Breast cancer transpires as one of the main source of deathly diseases among ladies around the world. Nevertheless, there is confirmation that early recognition and treatment can raise 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 utilizing 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 outcomes demonstrate that the proposed IDSS outperforms other systems in terms of diagnosis time and accuracy. The proposed system produces 99.94% classification accuracy. In addition, the proposed system reduces the required memory space.

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

I. INTRODUCTION

Breast cancer is one of the communal cancer-related diseases ladies around the world. Invasive breast cancer happens in nearly one out of eight (12%) women during their lifetime. On January 2012, greater than 2.9 million US ladies with breast cancer were alive [1]. Some of these ladies 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 ladies [2]. In 2015, the American Cancer Society's estimates that about 40,290 ladies will die from breast cancer and about 231,840 new cases of invasive breast cancer will be diagnosed in ladies, 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) [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.

The conventional system for diagnosing the disease depends on human skills to recognize 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 systems have been developed and functional in order to support specialists in the determination of the diagnosis decision process. 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 [6]. Other approaches are Artificial Neural Networks (ANNs), Bayesian Networks (BNs), Support Vector Machines (SVMs) and Decision Trees (DTs) [7].

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 [8], hybrid information gain and genetic filter/wrapper algorithm [9], multiple PCA with sparsity [10], decision rules (feature gene pairs) mining algorithm [11], genetic algorithm [12], discrete wavelet [13], mutual information, and various clustering techniques such as K-means clustering [14].

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 II presents the literature survey of related work while Section III gives an overview of the information entropy and the information gain. Section IV presents the proposed intelligent decision support system and describes the workflow of the proposed system. Section V illustrates the experimental results while Section VI presents the conclusion and the future progress of the research work.

978-1-4673-9652-3/16/$31.00 ©2016 IEEE

II. Related Work

Several techniques have been developed and tuned aiming to early detect breast cancer. In [15], a knowledge selection and classification of breast cancer disease using Adaptive Neuro-Fuzzy Inference System (ANFIS) is developed. The results show that the performance is improved and the accuracy of classification is 98.25%. Intelligent system that includes the artificial neural networks (ANN) based expert system for the automatic breast cancer diagnosis is becoming popular among researchers. In [16], 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 experimental results show that the accuracy obtained by applying the ensemble approach is better than that obtained by applying the individual approaches. In [17], 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. In [18], a medical decision support system is developed in light of Genetic Algorithm (GA) and Least Square Support Vector Machine (LS-SVM) for the detection diabetes on a Pima Indian Diabetes (PID) database. The suggested system outperforms of different standing systems which the outcomes display that based on the classification accuracy.

In [19], classification of cancer taking into account gene expression has given insight into conceivable treatment methodologies. Supervised learning systems that have been active to diagnoses cancers, a hybrid technique for feature selection in view of a feature selection method RelieIF and a genetic algorithm used to locate a set of features that can best distinguish between cancer subtypes or normal against cancer samples. Though, the k-nearest neighbor and linear SVM enhance the performance and execution of classification than different classifiers. In [20], 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.

III. 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. 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 [21]. Shannon [22] uses the concept of entropy in information processing and offered the concept of 'information entropy'.

If X is a discrete random variable with probability function P, its entropy is defined by:

$$H(X) = -\sum_{i} P(x_i)\ \log_2(P(x_i)) \qquad (1)$$

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:

$$H(C) = -\sum_{i=1}^{l} P(c_i)\ \log_2(P(c_i)) \qquad (2)$$

Here, $P(c_i)$ represents priori probability of the categorical variables C and c_i 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:

$$H(C) = -P(c_1)\ \log_2\big(P(c_1)\big) - P(c_2)\ \log_2\big(P(c_2)\big) \qquad (3)$$

For a gene X, it may have n possible values ($x_1, x_2, \ldots\ldots, x_n$). The corresponding conditional entropy is

$$H({}^{C}/_{X}) = -\sum_{j=1}^{n} P(x_j)\ \sum_{i=1}^{l} P(c_i\,/\,x_j)\ \log_2(P(c_i/\,x_j)) \qquad (4)$$

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, 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.

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

IV. 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, it selects significant features (feature selection) from the input patterns by using Information Gain (IG). Finally, the system employs two stages genetic algorithm, called Deep Genetic Algorithm (DGA), for data reduction as well as for breast cancer diagnosis.

A. Proposed System Workflow

Figure 2 shows the general framework of the proposed Intelligent Decision Support System (IDSS). 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.

B. Gene Microarray Dataset

The microarrays datasets are managed as a matrix with its rows denote the features (genes) and the columns denote the instances. Commonly, the whole number of genes considered just a minor number of gene expression data display a strong correlation. This implies that if a large number of genes studied; just a minor number show significant correlation with a certain phenotype. Therefore, feature selection is crucial for the classification process when analyze gene expression profiles validly [23].

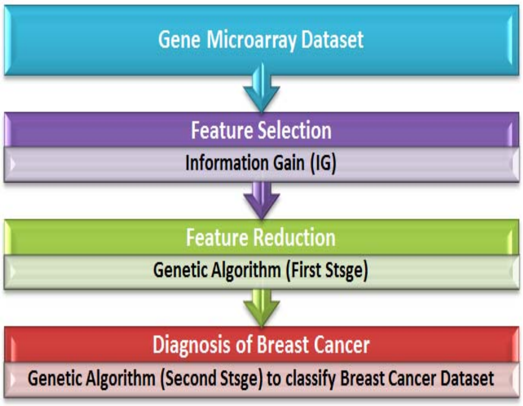

Gene Microarray Dataset
Feature Selection
Information Gain (IG)
Feature Reduction
Genetic Algorithm (First Stsge)
Diagnosis of Breast Cancer
Genetic Algorithm (Second Stsge) to classify Breast Cancer Dataset

Figure 1: Proposed IDSS Framework

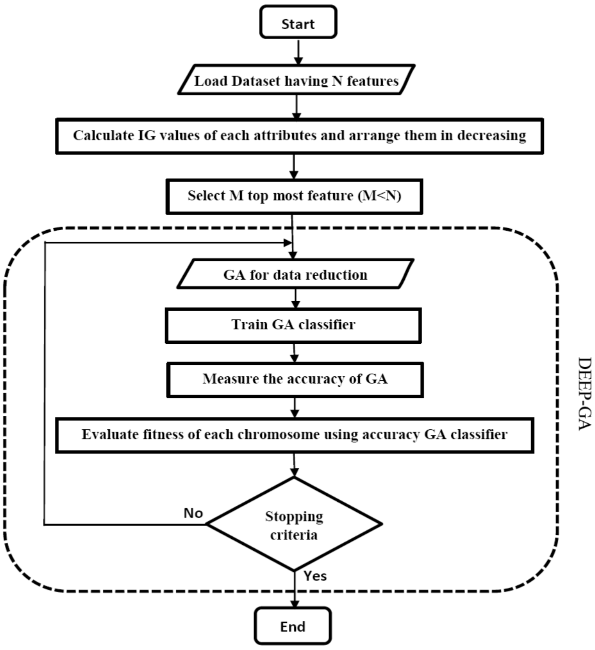

Start
Load Dataset having N features
Calculate IG values of each attributes and arrange them in decreasing
Select M top most feature (M<N)
GA for data reduction
Train GA classifier
Measure the accuracy of GA
Evaluate fitness of each chromosome using accuracy GA classifier
DEEP-GA
No
Stopping criteria
Yes
End

Figure 2: Workflow of the proposed system

C. 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 using the formula (2)
3) For each attribute (gene) in table 1, calculate the probability of all of its values. Compute conditional probabilities.
4) According to the probability obtained using the formula (4) for each gene (attribute), calculates conditional entropy.
5) Calculate Information Gain using classification attribute (5) for all genes (attributes).
6) Sort the outcomes obtained in step 5 and Select k Attribute with the highest gain as a compact subset of genes FS (depends on threshold).

The threshold plays an important role to form the number of features selected. It is controlled by the accuracy of classification. In effect, it has a direct influence of the performance of the DGA algorithm. 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.

Pseudo code for threshold selection procedure

```
Input:
minInfoGain, maxInfoGain
Output:
1st: list of possible suitable thresholds which are used for obtaining best features.
Begin
▸ initialize stepsCount to 1
      stepsCount ← 1
▸ get difference in integer between max and min
      firstStep ← 10*minInfoGain, lastStep ← 10*maxInfoGain
▸ update steps
      stepsCount ← lastStep – firstStep + 1
▸ get total count of steps
      stepValue ← └ (lastStep – firstStep) / stepsCount ┘
▸ set values
        min ← minInfoGain, max ← maxInfoGain, 1st ← empty list, 1st.add (min)
for  j←1 to stepsCount
        Min ← min + stepValue, 1st.add (min), 1st.add (max)
Return 1st
End
Input:
data: dataset loaded from data source, FLst: feature list that describe the data.
Classifier: the model classifier that produce the accuracy which cans be gained using the available data.
Output:
      SubFeature Lst: the reduced feature
Begin
     minIG ← InfoGain (data, FLst), maxIG ← InfoGain (data, FLst), Best acc ← 0
     1stThr ← getThrLst (minIG, maxIG), n ← 1stThr.count
for j←1 to n
     thr ← 1st.get (j),  features ← reduce (FLst, thr)
     model ← classifier.bulidModel (data, features, mode), acc ←model.classify (data)
if acc > bestAcc
     bestAcc ← acc, subFeatureLst ← features
End For
Return subFeatureLst
End
```

D. 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.

D.1 Genetic Algorithm based Feature Reduction (First Stage)

In the stage, the features chosen by IG are used for feature reduction by the genetic algorithm GA. 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.

a. Encoding and Initial Population

A straightforward encoding system to represent as much as possible of the available information, in which the chromosome is a series of bits whose length is dictated by the total number of genes. Each variable is associated with one bit in the series. On the off chance that the i^{th} bit is active (value 1), then the i^{th} gene is selected in the chromosome. Otherwise, a value of 0 shows that the corresponding features is ignored. Along these lines, each chromosome represents an alternate feature subset. Both, the active features and the number of them are produced randomly.

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

$$P_i = \frac{F_i}{\sum_i^{PopSize} F_i} \qquad (6)$$

Here F_i and PopSize are the fitness of ith 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 by the 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.
In GAs to date numerous types of crossover methodology have been attempted. In this study, a 2-point crossover operator was utilized, which picked two cutting points at random and on the other hand replicated single segments out of every parent.

d. Mutation

After that, all the chromosomes resulted from the crossover will go through a mutation operation and consequently, a new offspring is produced. . In the event that a mutation was available, both of the offsprings was mutated, and changed from 1 to 0, or from 0 to 1 as its binary representation 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 [24].

e. Fitness function

The fitness function evaluates every chromosome in the population with the goal that it might be ranked against the various chromosomes. The reason of the genetic search in the DGA approach is to look for ”good” feature subsets having the insignificant size and the maximum prediction accuracy. To accomplish this goal, we devise function of a fitness taking considering this criterion.

f. The principle of Setting Parameters of Genetic Algorithm

A genetic algorithm parameters setting is as follows:
(a) As the population size is too small, the premature convergence phenomenon is too big, make the fitness evaluation times increase sharply, the convergence speed significantly reduced. In this paper, the population size is 100.
(b) If the crossover probability is too low, may lead to a search block. If the crossover probability is too high, may lead to destruction of the original model. In this paper, the crossover probability is 0.8.
(c) If Mutation probability is too small, some gene bit premature loss of information may not be recovered. The rate of mutation was 0.2. Standard genetic operators, such as crossover and mutation, are applied without modification.

D.2 GA-Based Classifier (Second Stage)

D.2.1 Genetic Programming

In fact, a branch of genetic algorithm (GA) is GP, and the fundamental distinction between GA and GP is individual’s structure. GA individuals have string organized while GP’s individuals are trees [25]. The GP structure is as follows:

- **Produce introductory solutions for population:** To satisfy the population the initial solutions are made. Figure 3 shows the solutions in the population of GP.
- **Evaluate every solution by a function of fitness**: every solution is assessed to decide its fitness value.
- **By operators of genetic create a novel population:** The target of applying genetic operations on the population is to build the better quality population of the solutions. Reproduction, crossover, and mutation are genetic operators [26].
- **Reproduction:** In the next generation an amount of good solutions are chosen in view of their fitness value.
- **Crossover:** Parents are recombine parts from two great solutions, to make new solutions, titled “offspring” or “child”. Then two good solutions are picked. Corresponding to its fitness a solution being chosen. In GP, the sub-trees from parents are exchanged as shown in Figure 4.
- **Mutation:** The operator of mutation changes some piece of a solution randomly used to keep differing qualities in the population and to enable examination of distinct solutions. A location to be changed is chosen and a solution is chosen randomly [27]. In GP, a part is mutated by supplanting it with a chosen discretionary tree as shown in Figure 5.

Repeating these stages until the end criteria is met. The end measure for the run may be characterized by a most extreme number of generations or the best fitness quality.

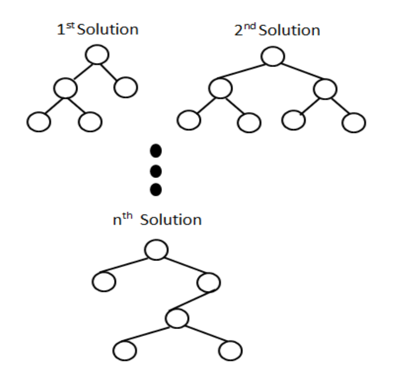


Figure 3: Solution Structure of GP

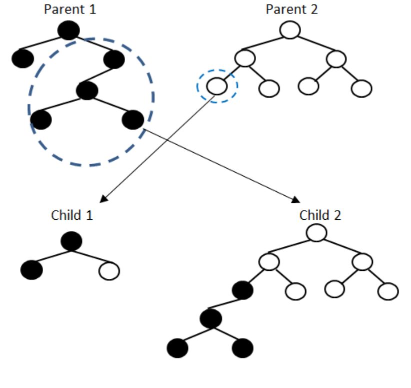


Figure 4: Crossover Operator in GP

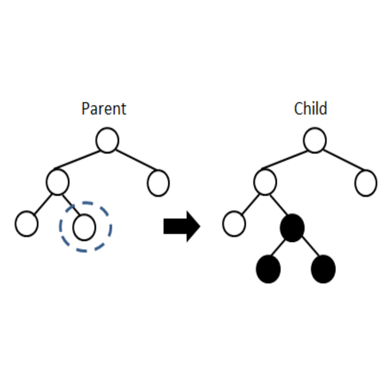
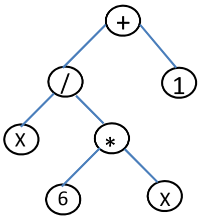


Figure 5: Mutation Operator in GP

D.2.2 Classification by Means of GP

GP based classifier is represented by a classification tree. The tree represents an arithmetic equation, as shown in Figure 6. The tree consists of symbols from the function series F and the terminal series T. The function series F comprises of arithmetic operators and the terminal set T comprises of 10 constants and variables defined as follows: F = {+, -, *, /} and T = {0.. 9, x_1... x_n}. The value of the expression level of genes is represented by variables.



$$\frac{x_9}{6 * x_4} + 1$$

Figure 6: Classification Tree and Corresponding Equation

The microarray dataset contains information for the variables (x_1… x_n). For an applicant to assess the fitness, its expression is assessed. On the off chance that the result of assessing an expression is more than 0, it is classified as Class 1. Otherwise it is classified as Class 2. An expression is assessed with data in the training series. The aggregate number of the right classification is considered as the fitness value of the expression.

- **Advantages and Disadvantages**

GP has been establishing to display a few focal points over other data-driven models (DDMs). Its significant favorable position is in its capability to create programs that can proficiently simulate complex procedures utilizing symbolic expressions [28]. Another point of preference of GP over other robust methods, for example, SVM is that it creates a straightforward and organized representation of the framework being demonstrated, without requiring from the earlier recognizable of the model structure. Be that as it may, in GP both the model structure and its parameters are being optimized, as they are both part of the search process. This gives GP the capacity to naturally recognize the data variables that contribute advantageously to the model and ignore those that don't, in this way reducing the dimensionality of the model. Additionally, GP advances models equipped for giving physical understanding into the input-output interactions inherent in demonstrated framework, rather than the SVM where difficulty still exists in extracting knowledge from the parameters. Then again, GP has its own particular restrictions. Basically, GP is not very powerful in discovering constants, and more importantly, it tends to create more complex functions as the forecast horizon growths [29].

V. Experimental Results and Discussion

The suggested system is evaluated by using the Skewed cancer gene expression database downloaded from the Kent Ridge Bio-medical Database website [30].

A. Microarray Datasets

The microarrays dataset is arranged as a matrix. The rows of the matrix represent the features (genes) while the columns represent the instances. The microarray gene expression data matrix split into two matrices; training data matrix and testing data matrix. For the feature selection step will be used the training data matrix and the result decreased train subset will train the implemented classifiers. For evaluating the proposed framework will be used the testing data matrix, by noting the number of test samples that the system will classify correctly. Detailed information about the microarray datasets summarizes in Table 1.

Table 1: Dataset Details

Datasets	Classes	Genes	Train Samples	Test samples
Breast Cancer	relapse, non-relapse (2)	24481	52 (27 relapse & 25 non-relapse)	26 (15 relapse & 11 non-relapse)

B. Performance Metrics

Table 2 summarizes the various performance metrics. The results are measured in contradiction of the following diagnostic performance measures. **P** positive instances and **N** negative instances True Positive (TP): the number of positive instances diagnosed correctly. True Negative (TN): the number of negative instances diagnosed correctly. False Positive (FP): the number of negative instances detected as positive (Type I error). False Negative (FN): the number of positive instances detected as negative (Type II error).

Table 2: Diagnostic performance measures Breast Cancer

	Total Population	**Condition Positive**	**Condition Negative**	**Total**
Predicted Condition	Predicted positive	True positive (TP)	False positive (FP)	(TP+FP)
	Predicted 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).

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

$$Sensitivity = \frac{\text{True positive (TP)}}{\text{(TP + FN)}} \tag{8}$$

$$\text{Specificity} = \frac{\text{True Negative (TN)}}{\text{(TN+FP)}} \tag{9}$$

C. Threshold Value

In the proposed framework, the first step uses IG for feature selection. IG value for each feature regulates whether this feature is to be selected or not. For testing the features the threshold value is practical, 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. A threshold for the results was recognized, after information gain values were considered for all features. After the computation process most papers demonstrate that most values of IG are zero, not many features have an effect on the category in a data set, signifying that these features are irrelevant for classification.

D. 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.

Together, the number of features and active features are produced randomly. We utilize the size of population of 100 individuals. Scattered crossover is taken randomly for every bit of the offspring, for joining parents of the preceding generation. The rate of crossover was established to 0.8. Roulette wheel as strategy selection and sampling uniformly was Applied, number of chromosomes which are taken in the next generation 50. Giving to that, consider a mutation operator with a probability rate of 0.2. An adjustment which incorporates mutating a random number of bits among the number of energetic features and 1 of the individual is presented. Fitness functions will be used; Classification accuracy (CA) of genetic algorithm as indicated by equation (7).

E. Experimental Results

Table 3 shows the experimental results of applying the proposed system on the breast cancer microarray gene expression dataset. The datasets are separated, with the first 66.67% (52 instances) of the instances being used for training, the next 33.33% (26 instances) for testing. The training set is used to train the DGA. Meanwhile, the testing set is used to obtain the test accuracy of the selected network. From the table, IG threshold value .2 is the optimal value for this dataset. At this value, features are reduced from 24481 attributes to 29 attributes in IG and reduced farther to 13 features by applying GA with 100 population size and 20 evaluation progress. In addition, the accuracy of classification is 99.94% by using the proposed system.

Table 3: Classification accuracy and extracted features under different IG threshold values for (19 instance)

IG Threshold Value	No. of features (Genes) After IG	No. of features (Genes) After GA	Accuracy of Classification
0.0	816	378	88.93 %
0.1	587	298	90.6%
0.2	29	13	99.94%
0.3	22	9	89.47%

In the current study, various thresholds are tested. For each threshold, the feature is chosen when the information gain value of this feature was higher than the predefined threshold, if not, the feature was not chosen. From this study, the best threshold value for this dataset is 0.2, where it achieves accuracy of classification of 99.94% .

The proposed algorithm have a complexity time compared to original Genetic algorithm which costs $O(\mathrm{nmpg})$ rather IG-GA consumes $O(\,n\log n + \mathrm{nmpg})$ and the proposed IG-DGA, $O(\mathrm{nmpg}(\,n\log n + \mathrm{nmpg}))$ Where n represents the number of samples, m represents the dimension of the data sets, p represents the population size and g represents the number of generations. The main disadvantage of the proposed is its time complexity, it is computationally expensive.

The proposed algorithm is classified as deep genetic feature subset selection based on the use of pros of information gain (IG) technique. The proposed system based on deciding which attribute have major impact on the classification accuracy. As each attribute have fixed time evaluation, that is calculated by IG method, so the total time of evaluation is variable depending on the total count of attributes which are sub-selected. So that, the proposed system generates the most optimal features set against others techniques used for reduction so the proposed have optimal execution time. Based on the experimental results, the proposed system is checked as the optimal disease time, in which time is computed based on evaluating each disease case using the interested set of attributes; same reduced before, and every attribute checking consumes constant cost by checking value of the attribute using decision statement; if or if-else. In turn, total disease time is the total time of checking all attributes which is minimized due to number of attributes are minimized as in the proposed compared against other methods (13 attributes). Also, Memory space occupied by redundant and unessential attributes are removed and hence lot of memory space is reduced.

GP strategy results with numerous different depths in balanced and unbalanced trees. The max depth of each individual is restricted to 3, and each nonterminal is forced to have exactly three children, which can be terminals or non- terminals. The size of population is 100, and the extreme number of generations is 20. The rate of crossover is 0.8, and the rate of mutation is 0.4. The fitness function for each individual is its accuracy on the validation set. The implementation of GP is based on the Pyevolve library [31]. Table 4 shows the classification accuracy, sensitivity and specificity of the proposed system (IG-DGA).

Table 4: Performance of Proposed Framework (sensitivity, specificity and accuracy)

Performance Metrics	Proposed Framework
Sensitivity (%)	99.47
Specificity (%)	97.93
Accuracy (%)	99.94

Figure 7 shows the classification accuracy of the proposed framework (IG-DGA) and 4 different algorithms (GA-LDA, GA-SVM, GA-NB and GA- C-MANTEC) reported in [32]. In the GA-LDA, GA-SVM, GA-BN, and GA-C-MANTEC, the feature selection method is Genetic Algorithm (GA) while the classifiers are Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Naive Bayes (NB) and the constructive neural network proposed (C-MANTEC) and all algorithm's use the same dataset and sample size that the proposed system utilize. By comparing the experimental results, the proposed system improves the sample classification accuracy, as shown in Figure 7, where the accuracy rates close to 100%. The experimental results demonstrate that the proposed procedure can enhance the constancy of the selection results as well as the sample accuracy of classification. Figure 8 shows the number of genes selected by each algorithm. By comparing the experimental results, the number of selected genes by the proposed algorithm is reduced to13.

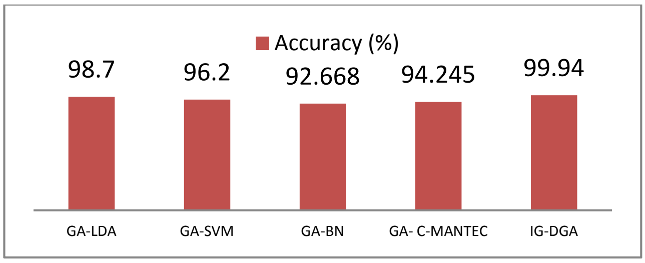


Figure 7: Classification accuracy of the proposed method and four other methods.

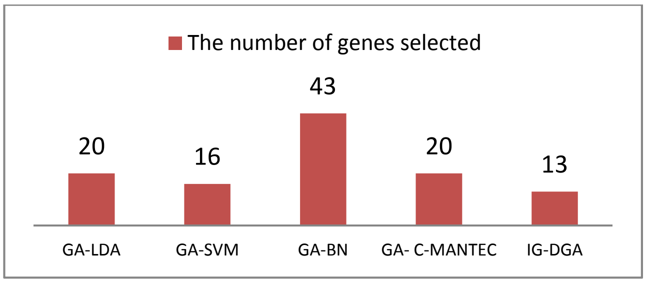


Figure 8: Number of genes selected by the proposed method and four other methods.

VI. Conclusions and Future Work

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, pre-select features are completed by IG however the DGA is utilized to additional distinguish a minor feature subset for increase accuracy of sample classification. A microarray dataset is utilized to assess the proposed algorithm. The experimental recommend that the proposed procedure can enhance the stability of the selection results as well as the sample classification accuracy. In terms of classification accuracy comparing with other methods, the proposed system can accomplish a good result also, decreasing medical errors, and minimizing life-threatening events caused by delayed or uninformed medical decisions. This algorithm achieves the optimum value for accuracy of classification percentage 99.94% from 24481 attributes that are reduced to 29 attributes in IG, which used 100 population size and 20 evaluation progress for GA feature Selection attributes reduced to 13. Memory space occupied by redundant and unessential attributes are removed and hence lot of memory space is reduced. In the future work, we will integrate various kinds of genomic data (interaction between gene expression profile and protein-protein dataset) to expand and upgrade the prediction accuracy when contrasted with utilizing gene expression a lone.

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 Hybrid GA-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] 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.

[7] 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.

[8] S. C. Shah and A. Kusiak,” Data mining and genetic algorithm based gene/SNP selection”, Artificial Intelligence in Medicine, Vol. 31, Issue 3, pp. 183-196, 2004.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] B. 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.

[16] 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.

[17] H. G. Zadeh, O. Pakdelazar, J. Haddadnia, G. R. Rad, and M. M. Zadeh, "Diagnosing Breast Cancer with the Aid of Fuzzy Logic Based on Data Mining of a Genetic Algorithm in Infrared Images", Middle East Journal of Cancer, Vol. 3, pp. 119-129, 2011.

[18] 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.

[19] H. Hijazi and C.Chan, "A Classification Framework Applied to Cancer Gene Expression Profiles", J Healthc Eng, NCBI, Vol. 4, pp. 255-283, 2013.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] D. A. Salem, R. A. AbulSeoud, 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.

[25] O. 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.

[26] 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**

[27] 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.

[28] A. Elshorbagy and I. El-Baroudy, "Investigating the capabilities of evolutionary data-driven techniques using the challenging estimation of soil moisture content," Journal of Hydroinformatics, vol. 11, pp. 237- 251, 2009.

[29] O. Giustolisi and D. Savic, "A symbolic data-driven technique based on evolutionary polynomial regression," Journal of Hydroinformatics, vol. 8, pp. 207-222, 2006.

[30] http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html. Accessed 26 May 2015.

[31] C. S. Perone, "Pyevolve: a python open-source framework for genetic algorithms," ACM SIGEVOlution, Vol. 4, no. 1, pp. 12–20, 2009.

[32] 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