Received: February 19, 2019

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Building an Ensemble Feature Selection Approach for Cancer Microarray Datasets Using Different Classifiers

Sabah Sayed[1] **Mohammad Nassef[1]*** **Amr Badr[1]** **Ibrahim Farag[1]**

[1]*Department of Computer Science, Faculty of Computers and Information, Cairo University, Egypt*
* Corresponding author's Email: m.nassef@fci-cu.edu.eg

Abstract: The challenge of processing the Microarray datasets with its high dimensionality opened multiple research directions. Different feature selection techniques have been employed to reduce the dimensionality of such Microarray datasets before being attempted by classification algorithms. This study presents an ensemble feature selection approach based on t-test and Genetic Algorithm with five different classification algorithms as its fitness function: Support Vector Machine, Random Forest, Nearest Centroid, K Nearest Neighbour, and Maximum Likelihood with 5-fold cross validation. The proposed approach has been applied on two different datasets for Lung cancer; Microarray Gene Expression and DNA methylation datasets aiming to find the Lung cancer biomarker genes. The experimental results showed that the three genes (DLX5, KRT5, and SELENBP1) resulted from processing both datasets have higher classification accuracy (92.31%) compared to separately processing the Gene Expression and the DNA methylation datasets with accuracies 90.38% and 86.54% respectively. Moreover, the classification accuracy achieved using the three aforementioned genes could not be achieved by other research studies unless by using more genes.

Keywords: Microarray gene expression, DNA methylation dataset, Lung cancer, Genetic algorithm, Feature selection.

1. Introduction

Cancer is a complex disease that results from abnormal biological processes. To understand these cancer processes, many measurement platforms have been developed and implemented in the field of bioinformatics [1]. One basic goal of the bioinformatics cancer systems is to infer the malignant drivers of those biological processes. Microarrays are one of the well-established tools used to identify and analyze the biological data. The main function of the Gene Expression Microarray experiments is to monitor the expression level of genes on the genome scale [2]. A Gene Expression matrix is the result of those experiments, where each row corresponds to a particular gene profile whereas each column represents the profile of an experimental condition.

DNA methylation (DNAm) is a common epigenetic mechanism, which controls the regulation of Gene Expression and is useful for early detection of cancer. Fortunately, DNAm Microarrays have been developed to measure the methylation level on the genome scale.

There are many databases, such as the Gene Expression Omnibus (GEO) and ArrayExpress, that serve as repositories of the resultant huge experimental data. Those databases contain data from Microarray experiments on a wide range of samples and under a variety of experimental conditions [3]. Moreover, the *International Cancer Genome Consortium* "ICGC" (http://icgc.org/) and the *The Cancer Genome Atlas* "TCGA" (http://cancergenome.nih.gov/) projects developed cancer-specific repositories that contain complete datasets related to many cancer types. For the cancer genome studies, those repositories are considered the main reference that offers the opportunity to test new computational approaches with real data [3].

The growing size of the biomedical data offers more research opportunities to analyze and discover new knowledge from this kind of data. Biomedical markers detection, diseases diagnosis, drug design

and classification of high-dimensional data are some of these research trends. Additionally, the high dimensionality is one of the main challenges in this biomedical data. More specific, a dataset consists of a small number of observations but with a large number of features that might be uninformative because they are either irrelevant or redundant [4]. In addition, the noise and variability of data add more complications.

From the large number of genes in Microarray Gene Expression dataset, only a small number of genes strongly correlate to the targeted disease. Many studies suggested that a small number of genes could form sufficient markers for a specific disease [5, 6]. Those few genes are usually called biomarker genes. Using only the biomarker genes in the classification phase not only reduces the computational effort but also increases the classification accuracy. A biomarker problem can be defined as selecting an effective and more representative gene subset. That being said, applying Feature Selection (FS) techniques in bioinformatics has become an important prerequisite step for model building. That's because most of the classification techniques were not designed to deal with huge number of irrelevant features. So, running them after FS techniques results in more efficient solutions [7].

Feature selection refers to selecting the most relevant features from the original feature space [4]. There are many FS techniques that differ in how each technique deals with the feature space to form a feature subset. In the classification problem context, these techniques can be divided into three categories: filter, wrapper, and embedded. The three categories differ in the way of combination between the feature selection search and the construction of the classification model. For more details about feature selection, the reader can refer to [7].

Obtaining a universally optimal feature subset requires using more than one FS technique [8]. For that, an ensemble FS approach runs different FS techniques where each technique produces a separate feature subset. Then, the ensemble FS approach combines the resulting feature subsets to form a final feature subset as its outcome. Ensemble FS approaches differ from each other in how they combine features. They may use averaging over multiple separate feature subsets [5, 9] that result from performing different runs of the same technique (for example, Genetic Algorithm) to assess the importance of each feature [10, 11], and using a collection of decision trees as random forest to assess the relevance of each feature [12, 13]. Ensemble FS approaches improve the robustness, stability, and generality but they require additional computations. The development of ensemble frameworks is a promising trend for improving the gene selection problem and the feature selection process in general because their flexibility and efficiency in dealing with high dimensional data [14].

Genetic Algorithm (GA) is one of evolutionary algorithms motivated by the biological theory of evolution and inspired by John Holland during the 1970s [15]. A GA implements the natural selection process by producing sets of solutions (population), each one called chromosome and represents a candidate solution for the underlying problem. The chromosome contains a group of features (genes). GA repeatedly produces solutions, calculates their fitness and terminates when the predetermined stopping criteria is met. The implementation of a GA is characterized by the fitness function and the genetic operators. The *fitness function* is used to assign a probability to all chromosomes in the population. This probability reflects the goodness of that chromosome and controls the reproduction process for the next generation.

The *genetic operators* are important to investigate the entire search space and to avoid the local minima. Crossover and mutation are the most popular operators. Crossover is used to swap genes between two randomly chosen chromosomes in one generation, producing two new chromosomes for the next generation. Crossover can be performed at single or multiple crossover points between chromosomes. It can be performed regardless the type of chromosome representation (binary or floating-point) [16]. A mutation means randomly flipping one or more gene in a chromosome according to a predetermined probability. The mutation process guarantees investigating all the search space of the underlying problem by altering gene values and so causing variations in the resultant solutions. There are many types of mutation; Binary Encoding mutation, Value Encoding mutation and Permutation Encoding mutation [17]. Elitism is a GA operator which allows keeping chromosomes with high fitness values to the next generation. A predefined probability is required to implement the elitism for number of generations.

Support Vector Machine (SVM) is a supervised classification algorithm developed by Cortes & Vapnik [18]. SVM was originally presented for binary classification problems, after that several modifications of SVM have been proposed to deal with multiclass problems. SVM generates a high dimensional feature space based on the attributes of features in the original data. Then, it tries to define a hyperplane or boundary to divide the feature space into different parts where each part represents the data points of one class [19].

The *Maximum Likelihood* (MLHD) algorithm finds the parameters' values for a given statistic that make the known likelihood distribution a maximum. MLHD has fundamental importance in the theory of inference and it is a basis of many other techniques in statistics [20]. The likelihood for a set of data is the probability of obtaining that particular set of data, given the chosen probability distribution model. The values of the model parameters that maximize the sample likelihood are called the Maximum Likelihood Estimates.

The *K Nearest Neighbour* (KNN) algorithm is considered from the simplest and most commonly used algorithm for classification, estimation and prediction. In KNN, to classify a new sample, find its K nearest neighbours from the dataset. The key is how to calculate the distance and how to choose the features to be used in the calculations. It is assumed that all features have the same impact on the distance. The most commonly used distance measurements are Euclidean Distance, Minkowski Distance and Mahalanobis Distance [21].

The *Nearest Centroid* (Nearcent) classifier is one of the simplest and powerful classifiers. It has been shown to perform well with Gene Expression Microarrays [22]. The centroid for a set of samples is the mean or median value of features in these samples. Nearcent simply classifies an unknown sample to the class with training samples whose mean (centroid) is closest to this sample. So the nearest centroid for an unknown sample is the centroid with the minimum Euclidean distance.

Random Forest (RF) [23] is a predictive model which is based on the classification trees. It uses an ensemble of decision trees to build a classification rule. By considering all trees in the forest, the final prediction is done by using a maximum vote scheme.

Various research studies have been attempted to apply different feature selection techniques over Microarray data with various goals. Using five-fold cross-validation, Abusamra [4] compared the classification performance of different feature selection and classification methods on the Gene Expression data of Glioma. Multiple feature selection methods have been used with three classification algorithms; SVM, KNN, and RF. By using only Gene Expression datasets, the results showed that combining FS methods with classification algorithms improved the classification accuracy by using fewer genes. Because of the relatively low resultant accuracies, Abusamra concluded that it is better to work by wrapper methods that integrate feature selection and classification methods to select better features and to have higher accuracy.

An ensemble-based feature selection technique was proposed in [24] to classify the Lung cancer subtypes based on DNAm data only. This technique produces three feature subsets from three separate methods (*Multi-category Receiver Operating Characteristic* (Multi-ROC), RF, *Maximum Relevance and Minimum Redundancy* (mRMR)). It then runs the *Incremental Feature Selection* (IFS) using *multi-class support vector machines* (Multi-SVMs) over a subset of the features overlapped between these feature subsets. More specific, although the common features were 45 in total, they used just 16 features that resulted in accuracy 84% which would be improved in case of using all the 45 features.

In [25], a comparative study between GA with *Constructive Neural Networks* and the classical *Stepwise Forward Selection* (SFS) algorithm in predicting the cancer outcome is conducted. The *Welch t-test* filtering method is embedded into the two algorithms. Those two algorithms have been applied on six cancer Gene Expression datasets. The results showed that the accuracy of SFS has not been improved.

A research was conducted by Garcıa [26] to analyze the effect of high-dimensional data on the classification of Gene Expression datasets. The *Gain Ratio* and *ReliefF* were used as gene ranking methods with six classifiers on four Gene Expression Microarray datasets. The results showed that regardless of the used gene ranking algorithm and classifier, the highest classification performance was achieved by using very few genes. Garcıa also proved that SVM has superior performance in cancer classification problems.

A multi-stage feature selection (MSFS) approach was proposed in [27] to find the optimal CpG-sites from a Lung cancer DNAm dataset. The MSFS approach combined three different filter feature selection methods: *Fisher Criterion*, *t-test* and *Area Under ROC Curve* (AUC). Thereafter, it applied Genetic Algorithm as a wrapper feature selection with *SVM Recursive Feature Elimination* (SVM-RFE) as a fitness function. By using the IFS strategy, subsets of 24, 13 and 27 optimal CpG-sites have been selected as biomarkers for the Breast, Colon and Lung cancer datasets respectively. Although the results were promising, the processing overhead of MSFS was enormous mainly because of the clustering step that MSFS begins with. This is in addition to using considerable CPG-sites in the classification process.

A Nested Genetic Algorithm (NestedGA) was proposed in [28] to define the biomarker gene subset by utilizing two types of data; Microarray Gene

Expression data and DNAm data. NestedGA consisted of two Genetic Algorithms (GAs); *outerGA* with SVM fitness function and *innerGA* with *Neural Network* fitness. The *t-test* filter method has been used as a preprocessing step before running the NestedGA as a wrapper feature selection method. NestedGA was applied on Colon cancer data with the aim to reach a gene subset representing the biomarkers for Colon cancer. NestedGA was also used to differentiate between the Lung cancer subtypes. The results showed that NestedGA noticeably outperformed InnerGA and OuterGA. Although the results of NestedGA are interesting, it suffers from huge processing overhead that cannot be avoided by parallelization due to the dependency between the nested GAs.

Except NestedGA, the aforementioned research studies targeted only one type of Microarray data (either Gene Expression or DNAm data), neglecting the biological relationship between those types of data in getting robust biomarkers. In addition, these studies generally have low accuracies and a considerable number of selected features that need more scientific validation.

This study aims to overcome the aforementioned shortcomings by constructing an ensemble feature selection approach to determine biomarkers for cancer diagnosis by utilizing two different types of Microarray data; Gene Expression data and DNAm data. Experimentally, the approach has been applied on Lung cancer datasets to differentiate between the lung cancer subtypes. The proposed approach combines filter and wrapper feature selection techniques. The *t-test* is used as filter feature selection technique, whereas, GA is used as wrapper feature selection technique with five different algorithms (KNN, MLHD, SVM, Nearcent, RF) as its fitness function. At last, IFS is applied individually on the five GAs to result in five different feature subsets. The resultant feature subsets are combined, and ranks are assigned for the features to produce a feature pool by using feature pool generation technique. The aforementioned steps are applied on Gene Expression data and DNAm datasets resulting in Gene pool and CpG-site pool respectively. The mapping between the Gene Expression data and DNAm data is used to obtain the genes related to all the CpG-sites in the CpG-site pool forming a new mapped gene pool. Finally, the common genes from the two gene pools are used in classifying the Lung cancer subtypes.

The rest of this paper is organized as follows. Section 2 describes in detail the used datasets and the main components and steps of the proposed approach. In section 3, the experiments with their results are stated and discussed. Finally, section 4 concludes the paper.

Table 1. Lung cancer dataset description

Dataset Type	**Gene Expression data**	**DNA Methylation data**
Dataset Variables	17,813 genes	27,578 CpG-sites
Dataset Function	• 66 samples for 5-fold cross validation training • 122 samples for testing	• 300 samples for 5-fold cross validation training • 11 samples for testing
Sample Type	• LUAD: 33 samples • LUSC: 155 samples	• LUAD: 151 samples • LUSC: 160 samples

2. Materials and methods

2.1 Datasets

The results presented in the next section are based on a Lung cancer Gene Expression dataset downloaded from The Cancer Genome Atlas (TCGA) https://tcga-data.nci.nih.gov/tcga/ and a TCGA DNAm dataset based on the Illumina IHM27k platform. These Lung cancer datasets contain two different cancer subtypes; LUAD and LUSC. Table 1 shows more details of the used datasets.

2.2 The proposed approach

Fig. 1 shows the pipeline of the proposed approach. First, the data is preprocessed before applying feature selection. After that, feature filtering is applied using *t-test* to select a subset of the top ranked features. The filtered feature subset is then fed as an input to the five GAs with different fitness functions (MLHD, RF, KNN, Nearcent, and SVM). By running each GA N times, the features resulting from the N runs are ranked in descending order based on their frequency. Next, the top-ranked features are incrementally accumulated producing incremental subsets of features that are ready for evaluation using classification. Eventually, the five different GAs result in five different feature subsets that are combined by removing the redundant features to generate a ranked feature pool. This pipeline is applied separately on both the Gene Expression and the DNAm datasets. So, the input features for the pipeline in Fig. 1 are either genes that result in a ranked Gene Pool, or CpG-sites that result in a ranked CpG-site Pool.

Using the annotation data that annotates genes to CpG-sites is important to combine the ranked Gene pool and the ranked CpG-site Pool. Initially, genes are mapped to CpG-sites by using the *minfi* and *IlluminaHumanMethylation27kanno:ilmn12:hg19* R packages. Each gene can be mapped to *h (h = 0 : 50)* CpG-sites. After that, an annotation table is built to maintain the reverse relationship so that CpG-sites in the ranked CpG-site Pool can be mapped to their corresponding genes forming another Gene Pool. Finally, the intersection between the two Gene Pools is obtained as shown in Fig. 2.

2.2.1. The preprocessing step

In the Gene Expression dataset, genes that have missing values are removed which result in decreasing the genes from 17,813 to 17,504. Similarly, for the DNAm dataset, CpG-sites decreased from 27,578 to 24,396.

2.2.2. Filter feature selection

Within the huge Gene Expression data, there are hundreds of genes that are redundant or irrelevant to the diagnosis of the targeted disease. So, it is important to reduce the number of genes in order to get good accuracy by the classification process. The *Student's t-test* is one of the most successful filter feature selection methods in terms of the quality of the ranked features [29]. The *Student's t-test* is applied on the two datasets using the *t.test()* R function as follows:

1. Divide samples into two classes; normal and tumour.
2. Calculate *p*-value for each feature reflecting how this feature is effective in separating classes.
3. Sort all the features according to their *p*-value ascending.
4. Select the best features (with lowest *p*-value).

For the Gene Expression dataset, the first 3,000 gene are selected, whereas, in the DNAm dataset the first 10,000 CpG-sites are selected.

2.2.3. Wrapper feature selection (GA)

A simple GA starts with initializing a population and running multiple iterations. Each iteration consists of some steps, which are known as GA operators (selection, crossover and mutation). At the end of each iteration, a new generation is created to be entered to the next iteration. The algorithm terminates when reaching the maximum number of iterations or finding the best solution. The flowchart of the GA algorithm is depicted in Fig. 3.

GA Chromosome Structure:

A chromosome *ch* with *n* features is represented as $ch = (g_1, g_2, \dots, g_n)$. These *n* features are randomly selected from the reduced feature set *F* produced from the previous stage. Each feature g_i is represented as an integer value that refers to the index of this feature in *F*. The chromosome structure is shown in Fig. 4.

The Steps of GA are as follow:

1. Initialize the GA initial population p_i with *Y* chromosomes each contains *y* feature selected from the filtered features (*F*) produced from the previous stage. Each chromosome is represented as an array of *y* indices that refer to the selected features. In first iteration the chromosomes are randomly initialized. For iteration *i* (*i*= 2, ... , *maxIter*), the chromosomes are initialized by using the best chromosomes from previous iteration *i-1*.
2. Calculate the fitness value f_i for each chromosome in the current generation using the determined fitness function.
3. Check if the termination conditions have been reached. The algorithm terminates with two conditions; reaching a solution with a predefined fitness value or reaching a predetermined number of iterations. In this case the algorithm outputs the best solution (subset of features) which is the chromosome with the highest fitness value in the current generation. Otherwise continue with the following steps.
4. To improve the performance, the best chromosomes in the current generation are selected to be persisted in the next generation with no change (elitism mechanism). To avoid trapping in local peaks, elitism is chosen to be performed for 9 consecutive generations and to be cancelled for the 10^{th} generation, this is repeated for all generations.
5. Apply Roulette Wheel Selection [30] to select subset *W* for crossover with length l_c. Steps of selection are as the following:
   (a) Generate random number *r* between 0 and sum of fitness values.
   (b) For each chromosome in the current generation, check if the chromosome's fitness is less than *r* then pick this chromosome to be in *W* and return to step a. Otherwise, check another chromosome.
   (c) Repeat steps *a* and *b* till l_c chromosomes are selected.

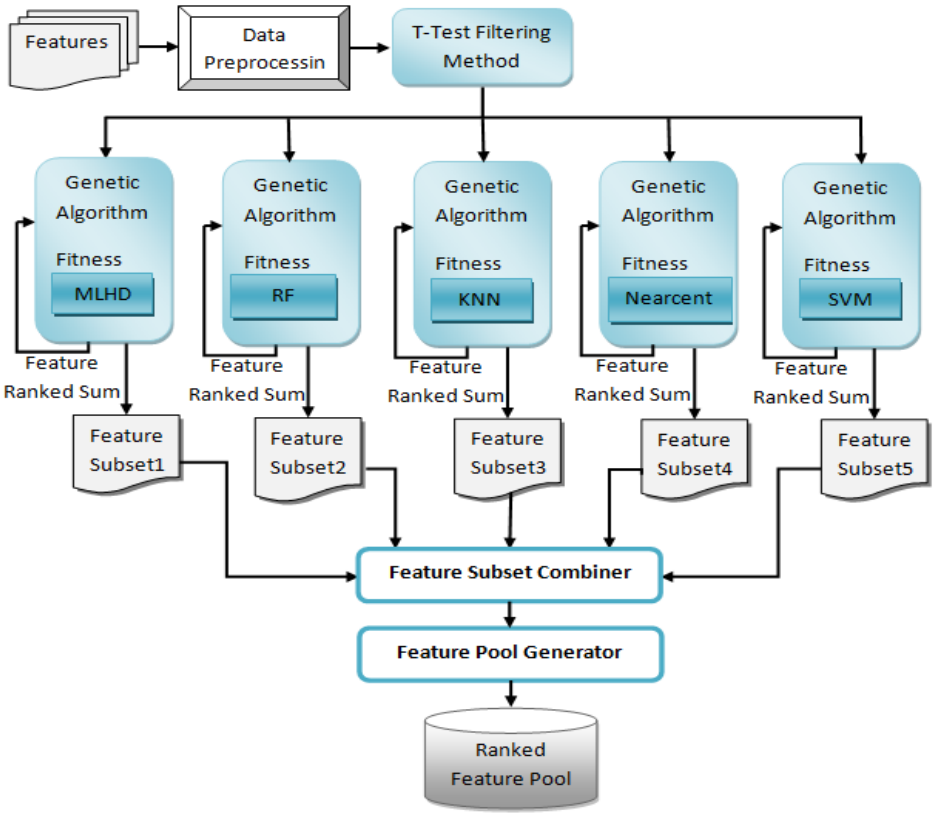

Figure. 1 Pipeline of the proposed approach

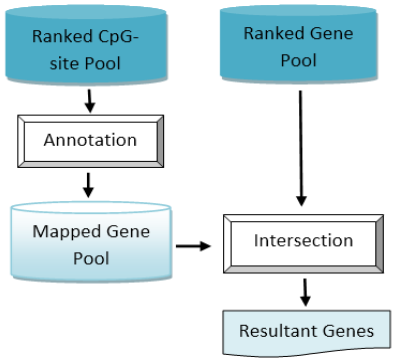

Figure. 2 Combining genes from CpG-site and Gene pools to get the intersecting genes.

6. Apply crossover by randomly selecting two parent chromosomes to create two new chromosomes. Crossover is applied as in Fig. 5.
7. Randomly select set of chromosomes with length l_m for mutation with mutation rate P_m. Perform a random single point mutation on these chromosomes by altering their genes values to ensure that a sufficient portion of the parameter space is explored.
8. Generate new generation from combining all chromosomes produced from the elitism, crossover, and mutation steps.
9. Replace old generation with the new one.
10. Repeat from step 2.

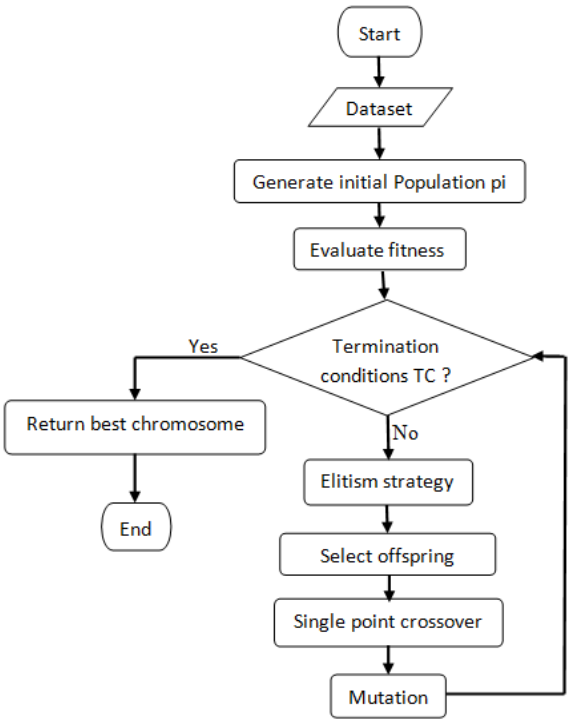

Figure. 3 GA flowchart

457	23	7098	. . .	5001	5

Figure. 4 Chromosome Structure. Each gene in the chromosome refers to an index of one feature

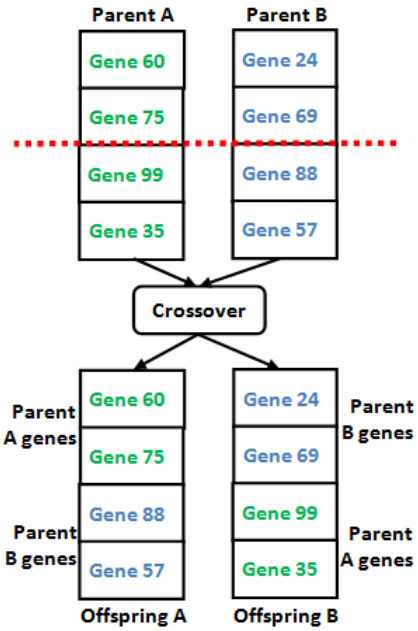

Figure. 5 The crossover mechanism

2.2.4. Feature sum ranker

In this step, *N* feature subsets are produced after running an individual GA N runs. The unique features are picked from those *N* subsets, and then sorted in descending order based on their cumulative frequencies over all the *N* subsets. The more frequent a feature, the higher its rank is.

2.2.5. Feature subsets combiner

A ranked features list *L* is Produced from the previous stage after repeatedly running an individual GA N runs. Next, IFS is applied to produce *S* feature subsets. The first feature subset is constructed from the first two top-ranked features. The remaining features are incrementally added one by one to produce new feature subsets. Each new subset is the same as its previous subset with a new feature added. Finally, *S* feature subsets are constructed where the i^{th} feature subset is: $sf_i = (f_1, f_2, \ldots, f_i)$ where $(2 \leq i \leq S)$. Applying the same process on the five different GAs results in five different feature subsets. Obtaining the intersected features between the five feature subsets is the second level of feature ranking.

2.2.6. Feature pool generator

Running the five different GAs in the previous step results in intersected feature lists. To generate a feature pool, four steps have been applied as follows:

1. Combine features from the intersected feature lists into a single list.
2. Remove the repetitions to produce the feature pool.
3. For each feature in the feature pool, compute the cumulative frequency over all intersected lists of features from GAs.
4. Sort feature pool according to the feature cumulative frequency.

Table 2. GA parameters and their description

Parameter	Description	Value
F	reduced features by *t-test* (genes, CpG-sites)	3,000 10,000
PSize	number of chromosomes in the population.	40
n	number of genes in a chromosome for Gene pool , CPG-site pool	30 50
maxIter	max number of iterations.	100
P_c	probability of crossover.	0.5
P_m	probability of mutation.	0.1
E	elitism selected chromosomes.	1
R	number of GA runs.	30
G	number of repeating GA runs.	10
goalF	required fitness value.	90 %

3. Results and discussion

The steps of applying the proposed approach are as follows. Firstly, the CpG-sites and Microarray genes with missing data are eliminated. Then, the *t-test* filtering method is applied (subsection 2.2.2) resulting in *F* feature set of top ranked Microarray genes. Five GA with fitness functions MLHD, Nearcent, SVM, KNN, and RF are performed on *F*. For each GA, feature sum ranker mentioned in subsection 2.2.4 is performed to get the near-optimal feature subset. This step is repeated *G* times for each GA producing *G* feature subsets. After that, feature subsets combiner is applied on those *G* feature subsets producing one feature list for each GA as in subsection 2.2.5. Finally, feature pool generator generates one pool from feature lists of the five GA. Table 2 shows the parameter settings for the proposed approach.

The Lung cancer datasets mentioned in subsection 2.1 is attempted by the proposed approach in two experiments. One experiment is for generating a Gene pool utilizing the Gene Expression Microarray dataset and the second one is for generating a CPG-site pool utilizing the DNAm dataset. The summary of the results of the two experiments is shown in Table 3. A third experiment is performed to get the intersected list of genes from both Gene pool and CPG-site pool. An annotation

Table 3. Summary of Gene pool and CpG-site pool results

Pool Type	Num of Features	Feature Frequency
Gene Pool	206 Genes	Genes with frequency 4 = 1 Genes with frequency 3 = 1 Genes with frequency 2 = 18 Genes with frequency 1 = 186
CpG-site Pool	194 CpG-Sites	CpG-sites with frequency 4 = 2 CpG-sites with frequency 3 = 4 CpG-sites with frequency 2= 25 CpG-sites with frequency 1=163

Table 4. The annotation of the three resultant genes to their corresponding CpG-sites according to HG19

Gene ID	Gene Name	Corresponding CpG-sites
1749	DLX5	cg00503840, cg01169726, cg02101486, cg06537230, cg06911084, cg08878323, cg09150117, cg11500797, cg11997899, cg12041387, cg13344740, cg13462129, cg16924616, cg18873386, cg20080624, cg24115040, cg27016494
33852	KRT5	cg04254916, cg23645091
8991	SELENBP1	cg18515587

step is required here to get the genes related to CpG-sites where each gene can be mapped to h ($h = 0 : 50$) CpG-sites. CpG-sites are mapped to genes by using the *minfi* and the *IlluminaIlluminaHumanMethylation27kanno:ilmn12 :hg19* R packages. The list of genes resulted from intersection is shown in Table 4.

Table 5 shows the accuracy for classifying the two Lung cancer subtypes for the Gene Expression dataset. The accuracy has been calculated three times; first using only the three genes shown in Table 4 which are resulting from the third experiment, and the second time using genes in Gene pool that have high frequency (from 2 to 5 genes). The last accuracy was calculated using the genes related to CpG-sites in CpG-site pool that have high frequency (from 2 to 5 CpG-sites). As shown in the table, the accuracy obtained by using the three resultant genes is higher than using up to five genes from the Gene pool or the CpG-site pool individually.

To justify the effectiveness of the proposed approach, it has been compared to the two latest approaches proposed in [24] and [28] applied on the same datasets for Lung cancer Gene Expression and

Table 5. The classification accuracies over the Gene Expression dataset using different gene subsets from different pools

Used Genes	Num. of genes	Accuracy
Gene pool & CpG-site pool	3	0.9231
Gene pool	2	0.8654
Gene pool	3	0.9038
Gene pool	4	0.8654
Gene pool	5	0.9038
CpG-site pool	2	0.8462
CpG-site pool	3	0.8654
CpG-site pool	4	0.9038
CpG-site pool	5	0.8846

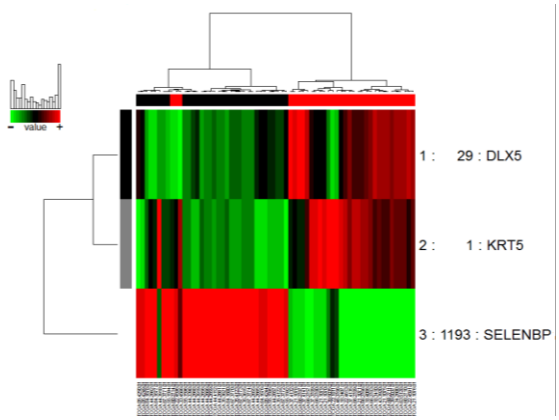

Figure. 6 The heatmap of the three genes intersecting between the Gene pool and the CPG-site pool

DNAm data. The accuracies are 92.31% for the proposed approach using 3 genes, 84.54% for the approach in [24] using 16 CpG-sites, and 87.60% for the NestedGA approach proposed in [28] using 3 genes. Moreover, the proposed approach is simple and fast compared to MSFS [27] and NestedGA [28] as it needs no complicated preprocessing steps.

As a step towards the validation of the resultant biomarkers, Fig. 6 depicts the heatmap generated for the three biomarkers genes (rows) with respect to the experimental samples (columns). It is clear from the heatmap that these three genes are cooperatively indicating high discrimination ability between the two Lung cancer subtypes samples. Gene *SELENBP1* has the highest discrimination ability, whereas gene *DLX5* has the lowest one.

3.1 Enrichment analysis

The three resultant genes have been validated with respect to the results published in previous researches. More specific, *DLX5* is one of the family

of Distal-less (DLX) *homeobox* genes. There are six family members, namely DLX1 to DLX6, where each gene has the same biological function for different species [31]. The authors in [32] reported that ERβ, like EGR3 and DLX5, is a direct positive regulator of NOTCH1 expression in keratinocytes and keratinocyte-derived squamous cell carcinoma (SCC) cells. As a result, this molecule (ERβ) is pointed to be a possible therapeutic target for differentiation therapy treatment of SCC.

According to [33] DLX plays an important role in tumour growth and progression because the deregulation of the DLX genes, including DLX5, was noticed in human solid tumours and hematologic malignancies. Moreover, DLX5 is reported to be an oncogene in Lymphomas and Lung cancers in [34, 35]. In Lung cancer cells, DLX5 overexpression is related to the tumour size and predictive of poor prognosis [36]. Furthermore, it is concluded in [37] that DLX5 is a target for the development of anticancer drugs and cancer vaccines, and it can act as a prognostic biomarker in clinic. In addition, DLX5 is considered a potential prognostic marker by [38-43].

In [44], *KRT5* is defined as one of the potential biomarkers for discriminating between LUAD and LUSC. More specific, KRT5 is suggested to have the highest diagnostic value for distinguishing between these two cancer types. KRT5 is also reported to be associated to Lung Cancer according to [45-49].

As shown in [50], the decreased *SELENBP1* is an early event in LUSC, and it can act as a novel potential biomarker for early detection of LUSC. Moreover, SELENBP1 is downregulated in many cancer types, such as Lung cancer according to [51-54].

4. Conclusion

In this study, an ensemble feature selection approach is introduced to find the biomarkers for classifying the Lung cancer subtypes using two different high dimensional datasets; Gene Expression and DNAm Microarray datasets. Five different feature selection techniques have been used as different fitness functions in five independent GAs. Then, the incremental feature selection strategy is applied to select the significant genes and CpG-sites resulting in Gene pool and CpG-site pool respectively. The resultant genes obtained by intersecting the Gene and CpG-site pools produced higher classification performance compared to using only genes from the Gene pool or CpG-sites from the CpG-site pool.

The main benefit of the proposed approach is to find the biomarker genes from two different types of cancer related data. The classification accuracy achieved by this study could not be achieved by other research studies unless by using more genes and higher computational complexity.

The resultant genes can be used to find more informative knowledge related to Lung cancer which can be utilized by future studies to discover more promising drugs. As a future work, the proposed approach can be applied on more than two types of data. Moreover, other optimization methods can be tried as a fitness function of Genetic Algorithm to get higher accuracies and other types of cancer can be treated.

References

[1] J. Gentle, W. Hardle, and Y. Mori, *Springer Handbooks of Computational Statistics*, 2010.

[2] G. Whitworth, "An introduction to microarray data analysis and visualization", *Methods in Enzymology*, Vol.470, pp.19–50, 2010.

[3] M. Vazquez, V. de la Torre, and A. Valencia, "Cancer genome analysis", *PLoS Computational Biology*, Vol.8, 2012.

[4] H. Abusamra, "A comparative study of feature selection and classification methods for gene expression data of glioma", *Procedia Computer Science*, Vol.23, pp.5–14, 2013.

[5] W. Li and Y. Yang, "How many genes are needed for a discriminant microarray data analysis", *Methods of Microarray Data Analysis*, pp.137–149, 2002.

[6] M. Xiong, X. Fang, and J. Zhao, "Biomarker identification by feature wrappers", *Genome Research*, Vol.11, pp.1878–1887, 2001.

[7] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics", *Bioinformatics*, Vol.23, pp.2507–2517, 2007.

[8] K. Yeung, R. Bumgarner, and A. Raftery, "Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data", *Bioinformatics*, Vol.21, pp.2394–2402, 2005.

[9] I. Levner, "Feature selection and nearest centroid classification for protein mass spectrometry", *BMC Bioinformatics*, Vol.6, No.68, 2005.

[10] L. Li, D. Umbach, P. Terry, and J. Taylor, "Application of the ga/knn method to seldi proteomics data", *Bioinformatics*, Vol.20, pp.1638–1640, 2004.

[11] L. Li, C. Weinberg, T. Darden, and L. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to

choose of parameters of the ga/knn method", *Bioinformatics*, Vol.17, pp.1131–1142, 2001.
[12] H. Jiang, Y. Deng, H. Chen, L. Tao, Q. Sha, J. Chen, C. Tsai, and S. Zhang, "Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes", *BMC Bioinformatics*, Vol.5, No.81, 2004.
[13] R. Dıaz-Uriarte and S. Alvarez, "Gene selection and classification of microarray data using random forest", *BMC Bioinformatics*, Vol.7, No.3, 2006.
[14] A. Jun, A. Mirzal, H. Haron, S. Member, H. Nuzly, and A. Hamed, "Supervised, Unsupervised and Semi-supervised Feature Selection: A Review on Gene Selection", *IEEE Transactions on Computational Biology and Bioinformatics*, Vol. 5963, pp.1-20, 2015.
[15] J. Holland, "Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence", *U Michigan Press*, 1975.
[16] D. Coley, "An introduction to genetic algorithms for scientists and engineers", *World Scientific Publishing Company*, 1999.
[17] R. Malhotra, N. Singh, and Y. Singh, "Genetic algorithms: Concepts, design for optimization of process controllers", *Computer and Information Science*, Vol.4, No.39, 2011.
[18] C. Cortes and V. Vapnik, "Support-vector networks", *Machine Learning*, Vol.20, pp.273-297, 1995.
[19] A. Karatzoglou, D. Meyer, and K. Hornik, "Support vector algorithm in R", *Journal of Statistical Software*, Vol.15, No.9 pp.1-28, 2006.
[20] I. J. Myung, "Tutorial on maximum likelihood estimation", *Journal of Mathematical Psychology*, Vol.47, pp.90-100, 2003.
[21] K. Venkatachalam and N. Karthikeyan, "Effective feature set selection and centroid classifier algorithm for web services discovery", *Indonesian Journal of Electrical Engineering and Computer Science*, Vol.5, pp.441-450, 2017.
[22] J. Won, J. Bok, M. Park, and S. Heun, "An extensive comparison of recent classification tools applied to microarray data", *Computational Statistics & Data Analysis*, Vol.48, pp.869-885, 2005.
[23] L. Breiman, "Random forests", *Machine Learning*, Vol.45, No.1, pp.5-32, 2001.
[24] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S. Ngai, and J. Shao, "Classification of lung cancer using ensemble-based feature selection and machine learning methods", *Molecular BioSystems*, Vol.11, pp.791-800, 2015.
[25] R. Luque-Baena, D. Urda, J. Subirats, L. Franco, and J. Jerez, "Analysis of cancer microarray data using constructive neural networks and genetic algorithms", In: *Proc. of the IWBBIO, International Work Conference on Bioinformatics and Biomedical Engineering*, pp. 55–63, 2013.
[26] V. Garcıa, J. Sanchez, L. Cleofas-Sanchez, H. Ochoa-Domınguez, and F. Lopez-Orozco, "An insight on the large g, small n problem in gene expression microarray classification", In: *Proc. of Iberian Conference on Pattern Recognition and Image Analysis*, pp. 483–490, 2017.
[27] A. Alkuhlani, M. Nassef, and I. Farag, "Multistage feature selection approach for high-dimensional cancer data", *Soft Computing*, Vol.21, pp. 6895–6906, 2017.
[28] S. Sayed, M. Nassef, A. Badr, and I. Farag, "A nested genetic algorithm for feature selection in high-dimensional cancer microarray datasets", *Expert Systems with Applications*, Vo.121, pp.233–243, 2019.
[29] E. Bonilla, B. Duval, and J. Hao, "A hybrid lda and genetic algorithm for gene selection and classification of microarray data", *Neurocomputing*, Vol.73, pp.2375–2383, 2010.
[30] K. Sastry, D. Goldberg, and G. Kendall, "Genetic Algorithms", *Springer US Boston MA, chap Genetic Algorithms*, pp. 97–125, 2005.
[31] G. Panganiban and J. Rubenstein, "Developmental functions of the distalless/dlx homeobox genes", *Development*, Vol.129, pp.4371–4386, 2002.
[32] Y. Sui, P. Ostano, S. Jo, J. Dai, S. Getsios, P. Dziunycz, G. Hofbauer, K. Cerveny, G. Chiorino, and K. Lefort, "Multifactorial erβ and notch1 control of squamous differentiation and cancer", *The Journal of Clinical Investigation*, Vol.124, pp.2260–2276, 2014.
[33] J. Li, P. Li, W. Zhao, R. Yang, S. Chen, Y. Bai, S. Dun, X. Chen, Y. Du, and Y. Wang, "Expression of long non-coding rna dlx6-as1 in lung adenocarcinoma", *Cancer Cell International*, Vol.15, No.1, pp.48, 2015.
[34] J. Xu and J. Testa, "Distal-less homeobox 5 (dlx5) promotes tumor cell proliferation by transcriptionally regulating myc", *Journal of Biological Chemistry*, 2009.
[35] X. Wang, Q. Xin, L. Li, J. Li, C. Zhang, R. Qiu, C. Qian, H. Zhao, Y. Liu, and S. Shan, "Exome sequencing reveals a heterozygous dlx5 mutation in a chinese family with autosomal-dominant split-hand/foot malformation", *European Journal of Human Genetics*, Vol.22, No.9, pp. 1105, 2014.

[36] T. Matsuo, M. Komatsu, T. Yoshimaru, K. Daizumoto, S. Sone, Y. Nishioka, and T. Katagiri, "Early growth response 4 is involved in cell proliferation of small cell lung cancer through transcriptional activation of its downstream genes", *PloS One*, Vol.9, No.11, 2014.

[37] T. Kato, N. Sato, A. Takano, M. Miyamoto, H. Nishimura, E. Tsuchiya, S. Kondo, Y. Nakamura, and Y. Daigo, "Activation of placenta-specific transcription factor distal-less homeobox 5 predicts clinical outcome in primary lung cancer patients", *Clinical Cancer Research*, Vol.14, No.8, pp.2363–2370, 2008.

[38] M. Morini, S. Astigiano, Y. Gitton, L. Emionite, V. Mirisola, G. Levi, and O. Barbieri, "Mutually exclusive expression of dlx2 and dlx5/6 is associated with the metastatic potential of the human breast cancer cell line mda-mb-231", *BMC Cancer*, Vol.10, No.1, pp.649, 2010.

[39] Y. Tan, R. Timakhov, M. Rao, D. Altomare, J. Xu, Z. Liu, Q. Gao, S. Jhanwar, A. Cristofano, and D. Wiest, "A novel recurrent chromosomal inversion implicates the homeobox gene dlx5 in t-cell lymphomas from lck-akt2 transgenic mice", *Cancer Research*, Vol.68, No.5, pp. 1296– 1302, 2008.

[40] N. Fujino, H. Kubo, T. Suzuki, C. Ota, A. Hegab, M. He, S. Suzuki, T. Suzuki, M. Yamada, and T. Kondo, "Isolation of alveolar epithelial type ii progenitor cells from adult human lungs", *Laboratory Investigation*, Vol.91, No.3, pp.363, 2011.

[41] Y. Tan, M. Cheung, J. Pei, C. Menges, A. Godwin, and J. Testa, "Upregulation of dlx5 promotes ovarian cancer cell proliferation by enhancing irs-2-akt signalling", *Cancer Research*, Vol.70, No.22, pp.1-10, 2010.

[42] B. Ricciuti, C. Mencaroni, L. Paglialunga, F. Paciullo, L. Crino, R. Chiari, and G. Metro, "Long noncoding rnas: new insights into non-small cell lung cancer biology, diagnosis and therapy", *Medical Oncology*, Vol.33, No.2, pp.18, 2016.

[43] E. Segal, N. Friedman, D. Koller, and A. Regev, "A module map showing conditional activity of expression modules in cancer", *Nature Genetics*, Vol.36, No.10, pp.1090, 2004.

[44] J. Xiao, X. Lu, X. Chen, Y. Zou, A. Liu, W. Li, B. He, S. He, and Q. Chen, "Eight potential biomarkers for distinguishing between lung adenocarcinoma and squamous cell carcinoma", *Oncotarget*, Vol.8, No.42, pp.71759–71771, 2017.

[45] T. Fukui, R. Shaykhiev, F. Agosto-Perez, J. Mezey, R. Downey, W. Travis, and R. Crystal, "Lung adenocarcinoma subtypes based on expression of human airway basal cell genes", *European Respiratory Journal*, Vol.42, pp.1332-1344, 2013.

[46] M. Ficial, C. Antonaglia, M. Chilosi, M. Santagiuliana, A. Tahseen, D. Confalonieri, L. Zandona, R. Bussani, and M. Confalonieri, "Keratin-14 expression in pneumocytes as a marker of lung regeneration/repair during diffuse alveolar damage", *American Journal of Respiratory and Critical Care Medicine*, Vol.189, No.9, pp.1142–1145, 2014.

[47] N. Smirnova, A. Schamberger, S. Nayakanti, R. Hatz, J. Behr, and O. Eickelberg, "Detection and quantification of epithelial progenitor cell populations in human healthy and ipf lungs", *Respiratory Research*, Vol.17, No.1, pp.83, 2016.

[48] A. Calio, A. Nottegar, E. Gilioli, E. Bria, S. Pilotto, U. Peretti,` S. Kinspergher, F. Simionato, S. Pedron, and S. Knuutila, "Alk/eml4 fusion gene may be found in pure squamous carcinoma of the lung", *Journal of Thoracic Oncology*, Vol.9, No.5, pp. 729–732, 2014.

[49] X. Xu, L. Huang, C. Futtner, B. Schwab, R. Rampersad, Y. Lu, T. Sporn, B. Hogan, and M. Onaitis, "The cell of origin and subtype of k-ras-induced lung tumors are modified by notch and sox2", *Genes & Development*, Vol.28, No.17, pp.1929–1939, 2014.

[50] G. Zeng, H. Yi, P. Zhang, X. Li, R. Hu, M. Li, C. Li, J. Qu, X. Deng, and Z. Xiao, "The function and significance of selenbp1 downregulation in human bronchial epithelial carcinogenic process", *PloS One*, Vol.8, No.8, 2013.

[51] N. Wang, Y. Chen, X. Yang, and Y. Jiang, "Selenium-binding protein 1 is associated with the degree of colorectal cancer differentiation and is regulated by histone modification", *Oncology Reports*, Vol.31, No.6, pp. 2506–2514, 2014.

[52] D. Caswell, C. Chuang, R. Ma, I. Winters, E. Snyder, and M. Winslow, "Tumor suppressor activity of selenbp1, a direct nkx2-1 target, in lung adenocarcinoma", *Molecular Cancer Research*, Vol.16, No.11, pp.1737–1749, 2018.

[53] M. Schott, M. de Jel, J. Engelmann, P. Renner, E. Geissler, A. Bosserhoff, and S. Kuphal, "Selenium-binding protein 1 is down-regulated in malignant melanoma", *Oncotarget*, Vol.9, No.12, pp.10445-10456, 2018.

[54] A. Pol, G. Renkema, A. Tangerman, E. Winkel, U. Engelke, A. Brouwer, K. Lloyd, R. Araiza, L.

Heuvel, and H. Omran, "Mutations in selenbp1, encoding a novel human methanethiol oxidase, cause extraoral halitosis", *Nature Genetics*, Vol.50, No.1, pp.120–129, 2018.