ISSN Print: 2394-7500
ISSN Online: 2394-5869
Impact Factor: 8.4
IJAR 2022; 8(7): 515-522
www.allresearchjournal.com
Received: 13-04-2022
Accepted: 15-06-2022

M Mohamed Divan Masood
Department of Computer Science and Engineering, School of Computer Applications, BS Abdur Rahman Crescent Institute of Science & Technology, Chennai, Tamil Nadu, India

TH Mohamed Ahadu Shareef
Department of Chemistry, The New College (Autonomous), Affiliated to the University of Madras, Chennai, Tamil Nadu, India

Kavipriya
Department of Computer Science and Engineering, Anna University, Chennai, Tamil Nadu, India

D Manjula
Department of Computer Science and Engineering, Anna University, Chennai, Tamil Nadu, India

Corresponding Author:
M Mohamed Divan Masood
Department of Computer Science and Engineering, School of Computer Applications, BS Abdur Rahman Crescent Institute of Science & Technology, Chennai, Tamil Nadu, India

Incorporating artificial bee colony optimization in gene based ranking system in microarray data

M Mohamed Divan Masood, TH Mohamed Ahadu Shareef, Kavipriya and D Manjula

DOI: https://doi.org/10.22271/allresearch.2022.v8.i7e.10010

Abstract
The Microarray data consists of various gene expression profiles which are used to identify disease, but the huge dimensions of the data make processing difficult. Assorted techniques like Artificial Bee colony are used to reduce the data and make its processing effective. The gene expression level is used to recognize diseases and help in their treatment. By identifying the early onset of disease, this technique will be able to arrest its further progress and control its prognosis, to a certain extent, through facilitating the development of appropriate medication using DNA datasets. The leukaemia dataset obtained the best accuracy at 81.667% with 35 genes in trial 3. The breast cancer dataset has the best accuracy of 70% with 20 genes in trial 1, and the prostate cancer dataset has its best accuracy at 85% with 35 genes in trial 2. Our proposed work focuses on the selection of affordable set of gene for a disease with an acceptable level of accuracy from the microarray dataset that is efficiently used to identify the cancer-causing genes in particular, or other disease-causing genes in general, in patients.

Keywords: Artificial bee colony, microarray dataset, support vector machine, dimensionality reduction

Introduction
Deoxyribonucleic acid (DNA) microarray data is used to analyses the given gene expression which describes the varied messenger ribonucleic acid (mRNA) in a cell. A microarray database is a depot comprising multifarious microarray gene expression data in huge dimensions, characterized by more clamor/noise than relevant data. In bioinformatics, gene expression profiling is used to analyze the structure of genes and identify disease by distinguishing normal cells from infected cells (Navarroa, *et al*. 2012) [18]. Microarray datasets constitute more data with clamor than with significant gene expressions, increasing the computation or processing time of the data, and culminating in adverse effects in terms of exploring drugs intended for a particular disease.
Classification techniques help classify diseases into certain groups associated with a particular disease, or simply into different relevant classes (Garro, *et al*. 2016) [10, 11]. Miscellaneous classification techniques are used to classify data correctly without classification errors and, further, develop a method that correctly classifies a DNA microarray. The genes so obtained best classify given disease with higher classification accuracy, which can be improved using different trials.
To reduce clamor in data, the filter method downsizes huge dimensional datasets through eliminating those gene expressions that do not suit the filter, thereby reducing the complexity and processing time of the data. Microarray classification is used to classify the given data into different sets of samples, with the samples forming a model. If the samples present in the data are fewer compared to the original data use of irrelevant samples, they negatively affect the testing of the model built, where the samples represent the volume of the genes. The huge volume of microarray datasets is aggregated with number of samples, which are less in size and it makes difficult to use sheer volume.
The filter method uses the Signal-to-Noise Ratio (SNR) to eliminate noise from the data by utilizing the mean and average values, given that the data is in Excel format. The mean and average for all the columns are computed separately, and if the value for the SNR is less than

The specified range, the corresponding columns are deleted and the size of the dataset reduced considerably. Filter methods are used to rank genes (Kar, *et al*. 2016) [15], which is the initial process. When the data is free of flaws, filter methods are applied and the genes ranked from the list generated from the filters. After ranking the genes, each filter classifies them using different classification methods. The best filter will be chosen, based on the accuracy percentage obtained, but the number of genes obtained will be smaller after classification. The issue is resolved by our proposed new method and uses only the SNR filter, which gives the highest accuracy.

In the dimensionality reduction process, after using the filter, we opt for the Artificial Bee Colony (ABC) algorithm to prune size of the data and it finds the best genes using bee behavior. The best valued informative genes are calculated and the top-ranked genes obtained as a result. It uses the greedy selection process and calculates probability values with the fitness function. Greedy selection picks the process with the minimum changes, while making changes that fit the criteria. It follows the heuristic method to make a local optimal choice that approximates to a global solution in a reasonable span of time.

Our proposed work of this paper is as follows. Section 2 describes the materials and methods of DNA microarray genomic datasets. Section 3 describes the result and discussion.

Materials and methods

Feature selection in microarray data

A microarray-planted gene expression outline has turned out to be an adequate procedure for the distribution, prediction, interpretation and medication of cancer. Constant changes in the progress of the disease have culminated in a tremendous quantity of data. Microarray data is best for the application of integrity and big data, which the size may vary over the time. The scrutiny of microarray datasets is crucial. They usually incorporate huge volumes of profiling, but only a portion of it constitutes genes that are quite distinctly expressed.

In feature selection framework the unnecessary feature is removed and it shrink the data dimensionality. Further, it impressive the accuracy of predictive system by doing away with insignificant aspect that produce flaws. There are three different types of feature selection approaches are developed with classification model namely filter technique, wrapper techniques and embedded techniques (Kar, *et al*. 2016) [15]. The filter technique uses the SNR filter to remove noise from the data by using mean and standard deviation. As far as the SNR is concerned, if the filtered result is below a particular range, the corresponding column is deleted and the genes reduced considerably using the filter. In the single filter method, the type of filter used decides the classification accuracy of the process.

The SNR filter gives the highest accuracy percentage and is used in the ranking process ((Ka, *et al*. 2015) [14]. Using the ranking process in the initial step will decrease the number of genes obtained. The t-test and other filters yield a lesser degree of accuracy when classification methods are applied. Different classification methods can be used for all filters to find the best filter of all, and the classification method, with the higher accuracy and number of genes is taken. After using the filter, the dimensionality reduction process in which genes are ranked is applied, increasing the chances of maximizing the number of genes. The obtained number of genes has every probability of leading to a particular disease, given the accuracy obtained in classification methods.

Dimensionality Reduction

Dimensionality reduction is executed to reduce the data considerably so as to curtail the number of genes, because the dataset contains more genes than samples. The ABC algorithm reduces the dimensions of the dataset (Kar *et al*. 2015) [14]. The ABC algorithm chooses the perfect set of genes using the greedy selection process and calculates probability values using the fitness function.

For both binary and multi-objective ranking, the ABC algorithm is used for feature selection (Li *et al*. 2016) [16]. Problems with both combination and optimization are dealt with in multi-objective ranking. The ABC methodology, works based on the bees 'behavior, is used because it finds the ranked genes most effectively. Dimensionality reduction is used to reduce the dataset by more than half, based on certain parameters or criteria which are applied for further processing in other modules. Reducing the data appropriately is vital, since they are used for further processing. Any negative values will affect the final result and the accuracy will, consequently, be diminished as a result. Correcting the data values in between is a difficult proposition, so it is crucial that the data be processed suitably to get the best results.

Ranking methods

Two supervised feature ranking methods are used. Ranking is done to arrange the data in order so the top-most genes, accorded the highest priority, that causes the disease can be easily ascertained. Ranking can be done using efficient feature ranking methods or using an algorithm such as the page rank. The data used here is simultaneously managed for overall boundary orientation as well as context boundary orientation. It mainly handles the adequate prospecting of high throughput data analysis (Liao, *et al*. 2015) [17]. Time complexity is very high. The biological significance of the gene subsets is further investigated. Pathway scrutiny and the gene regulatory network are all planted on a differential expression pattern. Therefore, employing our approach to this field and depicting its significance is the direction of our analytic process.

Ranking methods use either filter or wrapper methods. After the data is preprocessed, filters are used and the ranking of genes done. Occasionally the process may yield a huge list of genes, because the initial data may be large and applying a single filter turns up huge volume of genes are hard to rank since they will, more or less, have the same probability of pointing to a particular disease. Using ranking methods after filters gives fewer genes with a higher accuracy in disease identification. Ranking methods are always used for the final result because they arrange the genes that most likely cause the disease, as well as those that are least likely to cause it.

Robust and Efficient Feature Ranking (REFR) methods are used. Dimensionality reduction methods are typically categorized into two different types such as feature extraction and feature selection. The microarray data comprises the classic metric which aim to attain the throughput high and the few samples, where only take it to measure analytically (Peng, *et al*. 2016) [20]. The number of

samples present is small, making the process of ranking difficult. Ranking is efficient if the samples present are large and traditional data mining methods are used. The curse of dimensionality and problems with over fitting are binding in such situations. The implication of the data with the new representation is crucial. Ranking methods ascertain which genes are most likely to trigger the disease. Feature ranking is most applicable when its implications are emphasized. Feature ranking can be classified into the filter method, embedded method and wrapper method.

Classification Methods

The SVM is used in classification because it is a machine learning algorithm that classifies the given data more accurately than other classification algorithms (Alshamlana, *et al*. 2015) [2]. It is a machine learning algorithm used to classify the given data more accurately than any other classification algorithms. SVMs are supervised learning models and correlated learning algorithms that evaluate the data used for classification and regression analysis, and can perform non-linear classification as well. Consider a set of distinct training data, each belonging to either one or the other of two categories. An SVM training algorithm develops a model using the training data that allows new examples from one category or the other, making it a non-probabilistic binary linear classification. The accuracy of the training set is determined by giving the training set as a testing set and checking its accuracy.

The SVM uses the linear kernel, polynomial kernel and radial basis function, of which the linear and polynomial kernels yield the same results. In both cases, the obtained number of genes is lower and the radial basis function yields both a low number of genes and accuracy, so we consider only the linear and polynomial kernels. Other classification methods such as the K-nearest neighbor yield the best results in the T-test filter though the accuracy is lesser than with the SNR filter. The K-nearest neighbor used in the SNR filter gives slightly lesser accuracy, compared to the linear and polynomial (Alshamlana, *et al*. 2015) [2]; (Yang, *et al*. 2011) [26]. Neural networks can also be used, when obtained number of genes is higher, which is used to develop a model to train and test the huge amount of data.

Observations from the survey

It is obvious that the existing systems have certain shortcomings. They take into consideration only a single dataset, or a single filter is used for classification. Using several filters may reduce the number of genes and increase their accuracy in the classification, since fewer genes contribute most to the disease. The kernel functions used in classification methods are complicated, and simple classification techniques can be used instead and ranking is done in filter or ABC algorithm (Cho, *et al*. 2011) [8]; (Catto, *et al*. 2016) [7]; (Yu, *et al*. 2013).

The recent of data size of gene are insufficient and not needed to make advisory impact on the gene classification and selection procedure. In order to resolve these issues, the subset of diseases are used to reduce the dimension very tactically. The recent of reduced space many occupy the gene subset. These genes are used to diagnose disease and the most informative ones are selected from the microarray dataset (Huerta, *et al*. 2008) [13]; (Sasikala, *et al*. 2005) [22]. Microarray technology has become a revolutionary tool for understanding human disease. The resultant datasets seen in these cases are large and it is difficult to find informative genes from among them. Consequently, the list of genes is to be at a minimum and the efficiency maximum.

Results and discussion

The first step is to reduce the huge dimension of the data using methods such as feature selection and dimensionality reduction algorithms. The microarray technique offers the highest chances of revealing cancer-causing genes. Problems with dimensionality reduction create considerable difficulties in microarray data analysis, adversely affecting the selected gene. The benefits of the information obtained from a dataset helps estimate the degree of instability. The selection or eradication of consistent genes for cancer microarray datasets is a critical step in adequate classification.

Several machine learning methods have been research in the area of bioinformatics. These methods involve protracted computational time for the resolution and examination of large datasets on a traditional system with restricted computational capability. The large quantities of genes involved make the removal of irrelevant and redundant genes a challenge. The feature selection method uses the filter method with the SNR filter where the mean and standard deviation for all columns is calculated then SNR ratio likewise calculated using these two values. If the SNR value falls below a particular range, the entire column is deleted, else it is used. The noise level in the data must not exceed the range given. If it does not satisfy the condition, the result may be more irrelevant genes. The result obtained from the SNR filter contributes more to the disease than the deleted columns.

Dimensionality reduction uses the ABC algorithm to reduce the dataset and choose the perfect set of genes, applying the behavior of bees. ABC algorithm and ranking are included in module 2. Based on how bees search for food, genes are selected and the best found. The ABC algorithm is an excellent optimization technique based on bee behavior. Bees select their food using certain criteria such as choosing food sources nearby, opting for a food path based on the quantity of nectar available, and producing new solutions when the current solution can be improved no further, or when a condition called the limit or an abandonment criterion is reached. The algorithm gives different results for each trial because the cost function is used to choose the perfect set of genes.

The number of genes obtained in each trial also differs. The result is obtained as index values for genes, and the genes are ranked according to in the third module, the SVM classifier is used to classify the result obtained from the ABC algorithm and predict the accuracy of the datasets. The classification results obtained from the SVM classifier are more accurate than those from the others. Cross-validation is used in classification to split the data into training and testing sets and the data is trained using the training set and tested with the remaining data. The highest probability of their occurrence. Ranking finds the genes that contribute most to a particular disease.

The proposed system yields the best results when a ten-fold cross-validation is done. Both ten folds and five folds can be used in the validation. The ranked genes obtained from module 2 in the ABC algorithm are tested for accuracy. The original dataset is spit into training and testing sets, with the training set required to have more data than the testing set.

Accuracy can be improved by executing assorted trials in the ABC algorithm. When accuracy can be improved no further and continues to remain the same after a number of trials, it is fixed for the datasets in question.

This model proposes a system to choose perfect set of genes from the huge dimension of microarray data by reducing the dimensions of the dataset and comparing the data with three different datasets. It chooses the perfect or most accurate set of genes to analyze a disease with the highest classification accuracy. Various groups or classes of genes for a particular disease will be predicted.

The proposed system focuses on generating a list of genes, ranked using feature ranking methods, which have the highest probability of cancer. This study consists of two major stages:

Step 1: Dimensionality reduction of data to reduce the huge dimension of the microarray data, we initially apply a filter approach and an algorithm.

Step 2: Classification and ranking of genes. After dimensionality reduction, irrelevant genes from the microarray data are removed, as far as possible, and classification is done to check the accuracy of the data and rank the genes according to their position.

The final output of this system will be a list of the best genes, ranked according to feature ranking methods obtained after dimensionality reduction and classification. Data in the real world is spotted and cannot be used for tasks without cleaning and dimensionality reduction. Real-world data is often conflicting and inadequate, and is likely to incorporate flaws. Feature selection methods and the SNR filter are helps to eliminate irrelevant or noise data, and dimensionality reduction is performed where the ABC algorithm further reduces the data after the filter approach.

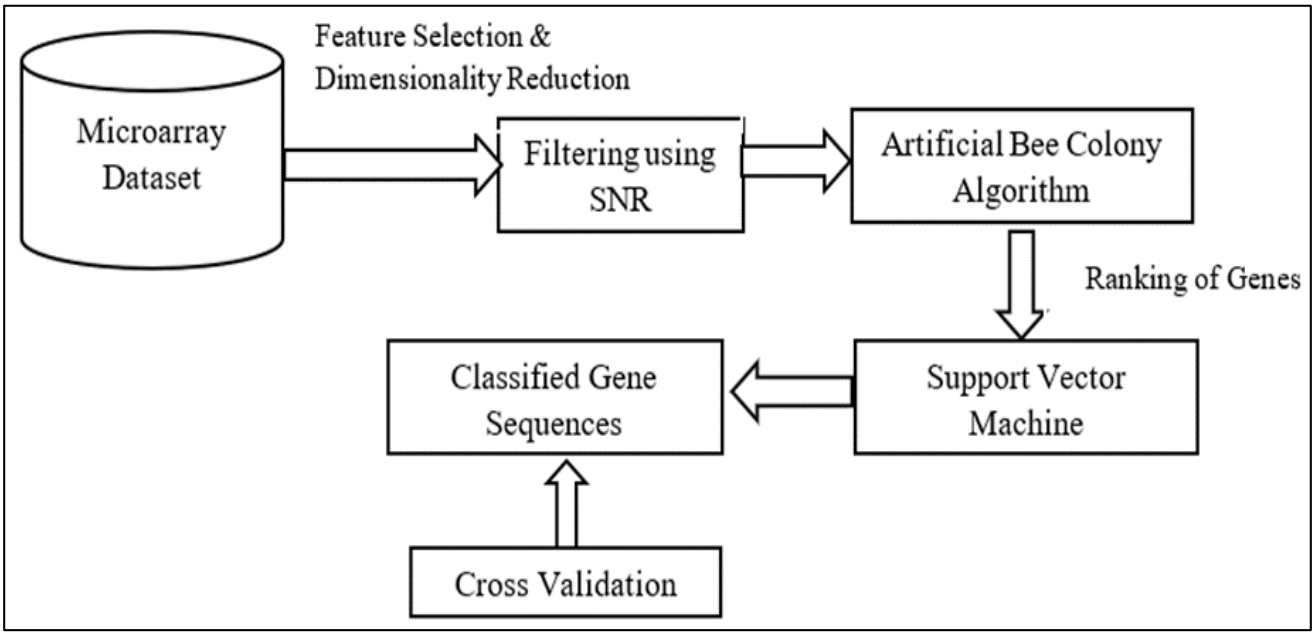

Fig 1: System Architecture for identified the classified genes

Filters are used to process the data to eliminate irrelevant genes and select the most informative genes from the microarray dataset. The SNR filter is used for the selection of genes because each filter method produces a different list of consistent genes. The idea is to choose a subset of informative genes from the filter method. Microarray data classification draws valuable information through machine learning and data mining techniques, and uses these processes to build a model to classify the given samples in miscellaneous categories (Peng *et al*. 2016) [20]. Conventional feature selection approaches are grouped into four categories: filter, wrapper, embedded and hybrid approaches (Yu, *et al*. 2012) [28]. Filter approaches assess each feature separately. The type of filter used decides the classification accuracy of the process.

Ranking selects the most informative genes from the list generated by the ABC algorithm. Each trial gives a different set of genes while each execution is performed. The ABC algorithm uses no special algorithm for ranking; instead, it chooses those genes which show the highest degree of accuracy. Ranking is also calculated using feature ranking methods, where goose with the lowest. This is done to select the most relevant gene that engenders disease.

Feature selection and dimensionality reduction

In recent year modern feature selection process is divided into four classes' namely wrapper, filter, hybrid approaches and embedded. The initial process of gene selection is done using filters. Filters are used to process the data to remove the irrelevant genes and select the informative genes from the microarray dataset. Usage of filters minimizes the initial gene, the search space and time complexity. Filters evaluate each feature separately and can be easily employed for high-dimensional data, by means of which the complexity can be lowered. Classification accuracy depends on the type of filter used. Microarray data uses machine learning in classification methods, and dimensionality reduction uses the ABC methodology to moderate the volume of the data, where the samples are built using the training data provided and tested using the tested set provided. If the column value is above the specified range, the column is taken. A single filter approach is used and classification accuracy is fully dependent on the filter.

Signal-to-Noise Ratio (SNR) Filter

The SNR filters reduce flaws in gene expression using standard deviation and mean values. The standard deviation and mean for every individual column are calculated, and the ratio between the two taken. If the value calculated is below the specified range, the corresponding column is deleted. SNR specifications are used in components such as amplifiers. The SNR filter consists of the following steps:

- The mean value for all the columns is calculated.
- The standard deviation for all the columns is calculated.
- The signal-to-noise ratio is calculated by dividing the mean by the standard deviation value.
- If the ratio is below 1, the corresponding columns (genes) are deleted, since they do not satisfy the condition stipulated by the filter.

Artificial Bee Colony (ABC) Algorithm

An excellent technique, which are expanded that is ABC from which the forging behavior of bees. Three different types of bees are used to accomplish convergence near the most favorable solution, i.e., the perfect set of genes.

Employed Bees: These laboring bees find their food source near the hive, and after finding the new path, replace the path of their old food source with the newly-found one. The greedy selection is done. The fitness function is used by the bees to evaluate the quality of each solution.

Onlooker Bees: Onlooker bees select the food source based on its availability near the hive. The quantity of food available must be plenty, depending on the quantity of nectar shown by each employed bee in which fit_i is the fitness value of the solution i and NB is the number of food sources that are equal to the number of employed bees.

Scout Bees: These bees search for new solutions, and when there are none, create random solutions. Where the solution can be improved no further, a condition called the "deadline" is attained,

Ranking of genes

Ranking is used to choose the perfect set of genes that has the highest classification accuracy in terms of the disease. The list generated from ranking gives both the best genes as well as those with the minimum probability of contributing to the disease, but we use only the best set of genes because it contributes most to the particular disease, while the minimum probability genes do not. We can use the minimum probability genes just to select those genes that do not cause disease. Ranking finds the genes with the highest probability of the disease. The steps for ranking are,

- The problem is initialized and defined, and the lower and upper bound set.
- ABC settings such as the maximum number of iterations, population size, and number of onlooker bees, abandonment limit parameter, and acceleration coefficient for the upper bound are defined.
- Population initialization is created using unfired (VarMin, VarMax, VarSize), and an empty bee structure and array to hold the best-cost values are created.
- The ABC algorithm-employed bee consists of an acceleration coefficient defined as ∅=a*unfired (-1, +1, VarSize) and the new bee position is calculated.
- The greedy selection process is applied.
- Bee position is calculated and the genes evaluated in onlooker bee phase.
- The fitness function is calculated using F(i) = Exp (pop(i). Cost/Mean Cost) and the probability value is also selected using P=F/sum (F).
- In the scout bee phase, the best solution is found and the best cost ever found is stored Best Cost (it) = Best Sol. Cost.
- An index is created to ascertain the position of the genes in the final result obtained.
- End.
- Start.
- For 1…n genes, do
- Find the position of each gene in the index.
- The most informative genes are obtained and those that contribute the most to a particular disease are ranked.
- Genes are ranked from the highest probability to the lowest.
- Ranks are generated for the genes obtained.
- End.

Classification of genes

Classification is done using the Support Vector Classifier, a machine learning algorithm that classifies the given data more accurately than any other. Machine learning is used in classification process because it develops programs that change, depending on whether the data is being exposed to Artificial Intelligence (AI). Classification splits the data into numerous classes, and the SVM classifier develops a model using the training sets and tests the model using the testing sets. The number of training sets used to train the model must be higher than the number of testing sets used.

Support vector machine

In this model, the SVM is used to find the mean accuracy of the process and perform cross-validation. The data set used here consists of only two classes: one that holds disease-causing cells, and the other comprising normal cells. The SVM consists of two or more classes, and the data we used has only two classes: one set representing cancer-causing genes and the other representing normal genes. Ten-fold cross-validation is used for evaluating the dataset. Accuracy is calculated for the data and the mean accuracy is taken to ascertain which data gives the best in terms of accuracy. The mean accuracy for all the three datasets is calculated. The SVM employs the following steps in the classification process:

- The data is loaded.
- The model is trained with the given training set and is tested using the testing set.
- Criteria, such as how many folds are given and the gamma value can be set manually.
- Classification is done and the model accuracy is established, by which means classification accuracy is determined.
- Accuracy can be improved by performing more trials.

Cross-validation

The genes obtained from the filtration process are evaluated to check the accuracy of the process and determine how informative the genes in question are. 10-fold cross-validation is performed in the classification algorithm. The steps for cross-validation include the following:

- Arrange the training sets in a random order.

- Divide the training example into k folds.
- For i=1…k
- Train the classifier using all examples that do not belong to fold i.
- Test the classifier on all examples in fold i.
- Computen$_i$, the numbers of wrongly classified examples in fold i.
- End.

Three datasets are used: leukemia, breast cancer and prostate cancer. The leukemia dataset contains around 7130 unique records, including class labels. Each record consists of a gene and its sample values. The leukemia dataset comprises two classes, ALL and AML, which are two types leukemia. It contains a total of 35 rows, of which 21 consist of ALL data and the remaining 14 of AML data, where both classes lead to Leukemia.

The breast cancer dataset consists of 16,386 unique attributes. Each record consists of a gene and its sample values. The breast cancer dataset consists of two classes, relapse and non-relapse. It contains a total of 20 rows, of which 13 are relapse (i.e., the sample which results in breast cancer) and the remaining 7 non-relapse, and not leading to breast cancer.

The prostate cancer dataset consists of 12,601 unique attributes, and each record consists of a gene and its sample values. The prostate cancer dataset consists of two classes, tumours and normal. It contains a total of 35 rows, of which 26 are tumours leading to cancer and the remaining 9 are normal, and not leading to cancer.

Table 1: Classification accuracy for the leukemia dataset

Trials	Accuracy	Number of genes
1	68.33	39
2	80	48
3	81.667	42
4	70	41
5	80.166	46

Table 1 shown below gives the accuracy and number of genes in the leukemia dataset. It shows the accuracy of the dataset, and the number of genes in each trial. The number of genes obtained varies for each trial; and, in some cases, all the trials contain the same number of genes. It is an evaluation metric used to measure the accuracy of the dataset, and the cross-validation is also done in the classification technique, which can be either five-fold or ten-fold. After the ABC algorithm is executed, the number of genes obtained for each trial varies, and each trial may give a different set of genes. Consequently, after each trial the accuracy of the genes is calculated using the support vector machine classifier. In this process, five trials are undertaken with every possibility of increased accuracy. When a certain criterion called the limit is obtained, wherein the accuracy of the datasets can be improved no further if more trials continue to be carried out, the trials are to be stopped. In the leukemia dataset, the best accuracy obtained is 81.667%.

Table 2: Classification accuracy for the breast cancer dataset

Trials	Accuracy	Number of Genes
1	70	41
2	60	39
3	68.166	48
4	69.667	40
5	69	41

Table 2 shown below gives the classification accuracy and number of genes in the breast cancer dataset. It shows the accuracy of the dataset and the number of genes each trial contains. The number of genes obtained varies for each trial, and in some cases, all the trials contain the same number of genes. After the ABC algorithm is performed, the number of genes obtained for each trial varies and each trial may give a different set of genes. After each trial, the accuracy of the genes is calculated using the support vector machine classifier. In this process, five trials are undertaken with every possibility of increased accuracy. When a certain criterion called the limit is obtained, wherein the accuracy of the datasets can be improved no further if more trials continue to be carried out, the trials are to be stopped. In the breast cancer dataset, the best accuracy obtained is 70%, which cannot be improved any further.

Table 3: Classification accuracy for the prostate cancer dataset

Trials	Accuracy	Number of Genes
1	83	41
2	85	39
3	76.66	35
4	84	34
5	81.667	37

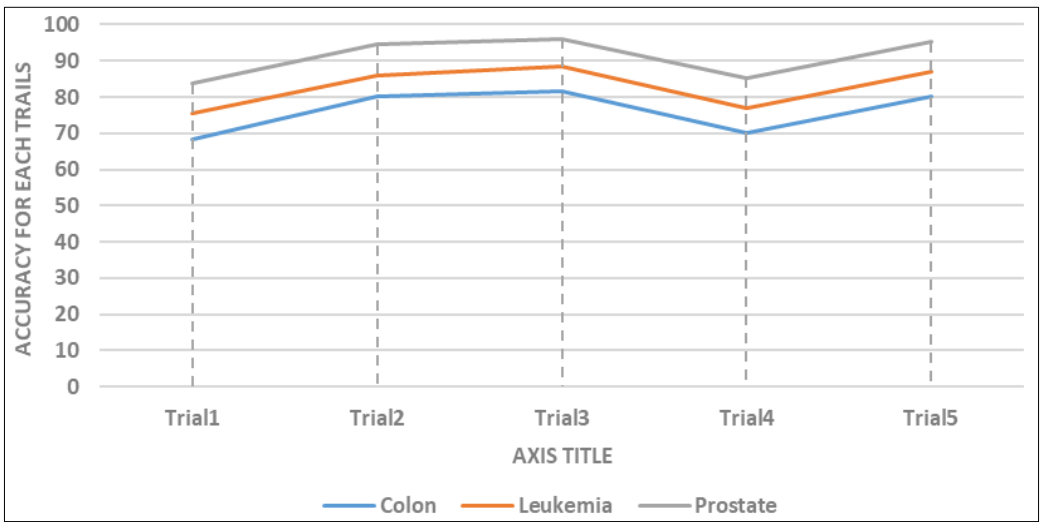

Fig 2: Classification accuracy for all the datasets in the five trials

Table 3 shown below gives the classification accuracy and number of genes in the prostate cancer dataset. It shows the accuracy of the dataset and the number of genes each trial contains. The number of genes obtained varies for each trial, and in some cases, all the trials contain the same number of genes. In this process, five trials are undertaken with every possibility of increased accuracy. When a certain criterion called the limit is obtained, wherein the accuracy of the datasets can be improved no further if more trials continue to be carried out, the trials are to be stopped. In the prostate cancer dataset, the best accuracy obtained is 85%, which cannot be improved any further.

Figure 2 shown below represents the accuracy of all the three datasets in the five trials. The x axis represents the number of trials while the y axis represents the accuracy of the three datasets obtained in each trial.

Figure 3 shown below gives the final accuracy of the genes. Five trials are performed for each dataset and the accuracy calculated using the classification technique. The best accuracy among the five trials is found using the graph above and they are plotted in a separate graph below. The x axis represents the number of datasets used - namely, leukemia, breast cancer and prostate cancer - and the y axis represent the accuracy of the best set of genes. The best accuracy obtained for each dataset is taken and a separate graph is plotted with the data. The x axis represents the number of datasets used - namely, leukemia, breast cancer and prostate cancer - and the y axis represent the accuracy of the datasets in each trial. Cross-validation is also performed in the classification method. The accuracy of the genes cannot be further improved since there is a limit set. The leukemia dataset obtained the best accuracy at 81.667% with 35 genes in trial 3. The breast cancer dataset has the best accuracy of 70% with 20 genes in trial 1, and the prostate cancer dataset has its best accuracy at 85% with 35 genes in trial 2. In phase 1, the best accuracy obtained is only 72% whereas in phase 2, it is increased by 10%. When a certain criterion called the limit is obtained, the accuracy of the genes can be improved no further. Any number of trials that follows thereafter gives the same accuracy, though the set of genes varies in each trial carried out.

Microarray datasets taken from various repositories are processed to identify the best gene subset that has the highest probability of a particular disease. First, the dimensions of the dataset are reduced. Following the dimensionality reduction process, the best set of genes obtained is classified and the classification accuracy determined. Thereafter, the data most predisposed to cancer, or any other disease, are sorted according to the order of the gene values, from the highest to the lowest. Only selected data from each process are analyzed, and the list of genes obtained is high. To further reduce the list and obtain the best set of genes from them, we rank the genes according to their position, using a single filter and a single classification method. Future work can focus on using multiple filters or using a range of classification methods. The data obtained from the filter is rather limited, so methods to improve the genes obtained are vital, and the dataset used to diagnose the disease has demonstrated the highest probability.

Conclusion and future work

In this present method has to identify the classified gene sequences using the microarray datasets. In our approaches has been used three different kinds of datasets namely leukemia, breast and prostate cancer datasets. Microarray datasets taken from various repositories are processed to identify the best gene subset that has the highest probability of a particular disease. First, the dimensions of the dataset are reduced. Following the dimensionality reduction process, the best set of genes obtained is classified and the classification accuracy determined. Thereafter, the data most predisposed to cancer, or any other disease, are sorted according to the order of the gene values, from the highest to the lowest. Only selected data from each process are analyzed, and the list of genes obtained is high. To further reduce the list and obtain the best set of genes from them, we rank the genes according to their position, using a single filter and a single classification method. Future work can focus on using multiple filters or using a range of classification methods. The data obtained from the filter is rather limited, so methods to improve the genes obtained are vital, and the dataset used to diagnose the disease has demonstrated the highest probability.

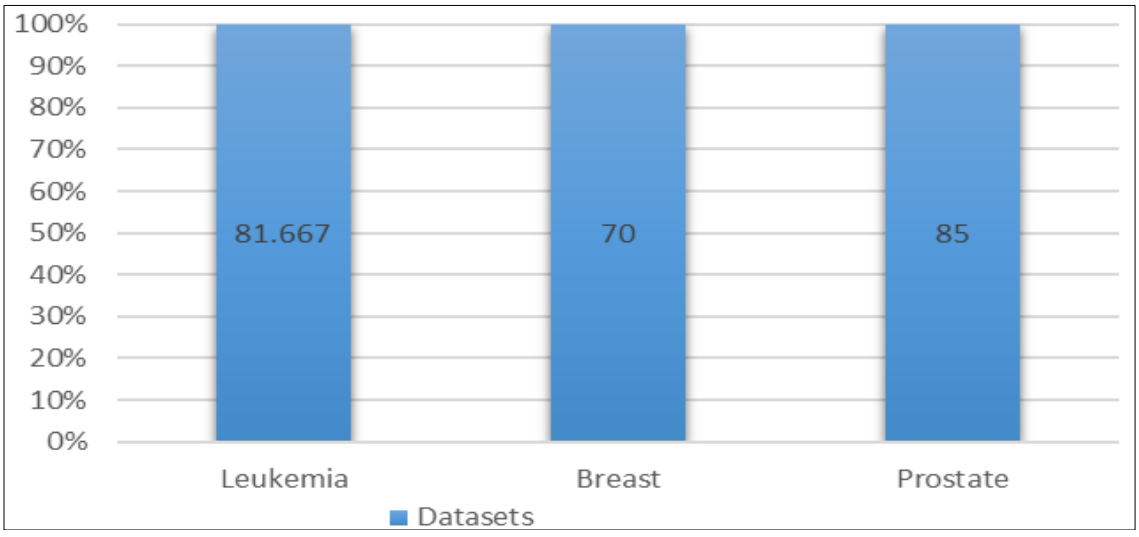

Fig 1: Classification accuracy of the topmost genes

Conflict of interest: The authors declare that there is no conflict of interests regarding the publication of this article.
Funding: Not applicable.

References

1. Aljawarneh SA, Jaradat R, Maatuk AM, Alhaj A. Gene Profile Classification: A proposed solution for predicting possible diseases and initial results. In engineering & MIS (ICEMIS), International Conference on IEEE, 2016, 1-7.
2. Alshamlana MH, Badra HG, Alohalia AY. Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Journal on Computational Biology and Chemistry. 2016, 49-60.
3. Babu M, Sarkar K. A comparative study of gene selection methods for cancer classification using microarray data. Second international conference on research in computational intelligence and communication networks (ICRCICN). 2016, 204-211.
4. Baena RML, Urda D, Subirats JL, Franco L, Jerez JM. Analysis of cancer microarray data using constructive neural networks and genetic algorithms. International Work-Conference on Bioinformatics and Biomedical Engineering, 2013, 55–63.
5. Bouazza SH, Zeroual A, Auhmani K. Gene expression data analyses for supervised prostate cancer classification based on feature subset selection combined with different classifiers. IEEE Conference, 2016, 163-168.
6. Canedo VB, Maroo NS, Betanzos AA. Distributed feature selection: an application to microarray data classification. Applied Soft Computing. 2015;30:136–150.
7. Catto JW, Abbod MF, Wild PJ, Linkens DA, Pilarsky D, Rehman I, *et al*. The application of artificial intelligence to microarray data: Identification of a novel gene signature to identify bladder cancer progression, 2016, 398–406.
8. Cho J, Kim D. Intelligent feature selection by bacterial foraging algorithm and information theory. Advanced Communication and Networking. 2011;199:238–244.
9. Etemadi R, Alkhateeb A, Rezaeian I, Rueda L. Identification of Discriminative Genes for Predicting Breast Cancer Subtypes. IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2016, 1184-1188.
10. Garro AB, Rodriguez K, Vazquez AR. Classification of DNA microarray using artificial neural networks and ABC algorithm. Journal on Applied Soft Computing, 2016, 548-560.
11. Garro AB, Rodriguez K, Vazquez AR. Generalized Neurons and its application in DNA Microarray Classification. IEEE Congress on Evolutionary Computation, 2016, 3110-3115.
12. Ghorai S, Mukherjee A, Sengupta S, Dutta PK. Cancer classification from gene expression data by NPPC ensemble. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2010;8:659–671.
13. Huerta EB, Duval B, Hao JK. Gene selection for microarray data by a LDA-based genetic algorithm. Pattern Recognition in Bioinformatics. 2008;5265:250–261.
14. Kar S, Sharma DK, Maitra M. A Comparative study on Gene ranking and classification methods using microarray gene expression profiles. Michael Faraday IET International Summit, 2015, 596-600.
15. Kar S, Sharma DK, Maitra M. Gene selection for tumor classification using resilient backpropagation neural network. IEEE Transaction, 2016, 5190-5195.
16. Li X, Li M, Yin M. Multi objective Ranking binary artificial bee colony for gene selection problems using microarray datasets. IEEE/CAA Journal of Automatic Sinica, 2016, 1-16.
17. Liao B, Jiang Y, Liang W, Peng L, Hanyurwimfura D, Li Z, *et al*. On efficient feature ranking methods for high-throughput data analysis. IEEE/ACM transactions on computational biology and bioinformatics. 2015;12(6):1374-1384.
18. Navarroa FF, Martíneza HC, Ruizc R, Riquelmeb CJ. Evolutionary generalized radial basis function neural networks for improving prediction accuracy in gene classification using feature selection. Journal on Applied Soft Computing. 2012, 1787-1800.
19. Nguyen T, Khosravi A, Creighton D, Nahavandi S. A novel aggregate gene selection method for microarray data classification. Pattern Recognition. 2015, 16–23.
20. Peng Q, Lv J, Chen X, Sun Z. A multi-objective heuristic algorithm for gene expression microarray data classification. Journal Expert systems with applications. 2016; 13-19.
21. Sahu B, Mishra D. A novel feature selection algorithm using particle swarm optimization for cancer microarray data. Procedia Engineering. 2012;(38):27-33.
22. Sasikala S, Balamurugan SA, Geetha S. A novel feature selection technique for improved survivability diagnosis of breast cancer. Procedia Computer Science. 2005;(50):16–23.
23. Sharbaf F, Mosafer S, Moattar M. A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization. Journal on Genomics. 2016, 231-238.
24. Shen Q, Shi WM, Kong W. Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. Computational Biology. 2008, 53–60.
25. Tabakhi S, Najafi A, Ranjbar R, Moradi P. Gene selection for microarray data classification using a novel ant colony method for optimization. Journal on Neurocomputing, 2015, 1024-1036.
26. Yang F, Mao KZ. Robust feature selection for microarray data based on multi criterion fusion. IEEE/ACM Transactions on Biology and Bioinformatics. 2011;8(4):1080–1092.
27. Yu H, Ni J, Zhao J. Acosampling: an ant colony optimization-based under sampling method for classifying imbalanced DNA microarray data. Journal on Neurocomputing. 2016;(101):309-318.
28. Yu L, Han Y, Berens ME. Stable gene selection from microarray data via sample weighting. IEEE/ACM Transactions on computational biology and bioinformatics. 2012;9(1):262–272.
29. Zhang SW, Huang DS, Wang S. A method of tumor classification based on wavelet packet transforms and neighborhood rough set. Journal on Computational Biology. 2010;(40):420–437.