GENETIC ALGORITHM WITH BAGGING FOR DNA CLASSIFICATION

Balamurugan E
Department of Computer and Mathematical Sciences,
University of Africa Toru-orua, Nigeria, Africa
rethinbs@gmail.com

Jackson Akpajaro
Department of Computer and Mathematical Sciences,
University of Africa Toru-orua, Nigeria, Africa
jakpojaro@yahoo.com

Submitted: *Aug, 14, 2021* ***Revised:*** *Oct, 21, 2021* ***Accepted:*** *Nov, 05, 2021*

Abstract: Accurate classification of cancer plays an important role for cancer treatment. The advancement of microarray technologies improves the accuracy of cancer diagnosis. Recently, scientists identify more informative genes from thousands of genes for accurate cancer detection. In this paper, Genetic Algorithm (GA) with bagging is developed for DeoxyriboNucleic Acid (DNA) classification. To remove the noise and data integrity, GA is applied to find the informative genes from the microarray data. It uses Backward Selection (BS), Forward Selection (FS) and Branch and Bound Selection (BBS) algorithms to select the sub-set of genes. Then bagging is employed to classify the selected genes to normal or abnormal. The evaluation of DNA classification system is performed on five cancers; colon, Central Nervous System (CNS), ovarian, leukemia and breast. Results show that the accuracy of GA-BBS with bagging algorithm is better than GA-BS and GA-FS with bagging. For all datasets, GA-BBS with bagging provides no misclassification and gives the highest performance (100%) in terms of sensitivity, accuracy and specificity. Based on results, it is concluded that 'best' prediction system is GA-BBS with bagging classifier.

***Keywords*:** DNA classification, genetic algorithm, feature selection, ensemble method, decision tree, bagging.

I. INTRODUCTION

The International agency for research on cancer estimates that the number of new cases in 2020 is 19.3 million. It is predicted that by 2040, number of new cases will increase to 30.2 million worldwide. A framework for gene selection for the diagnosis of cancer is discussed in [1]. The hybrid selection technique of GA balances the efficiency and accuracy using the obtained best sub-set of gene expression data. An approach for the classification of a DNA microarray using complex network is described in [2]. A structural algorithm enables the entry variables to be picked for distinct nodes. A hybrid technique that integrates the GA with Particle Swarm Optimization (PSO) is used for discovering optimal classification parameters.

The binary PSO wrapper technique is used in [3] to get the most relevant genes for the categorization. The early convergence by local stagnation problems never produces acceptable classification accuracy. The binary PSO based wrapper is checked using stratified 5-fold cross-validation by means of Naive-Bays (NB), k

Nearest Neighbour & Support Vector Machine (SVM) classifiers. The re-sampling and feature selection based approach for DNA classification is discussed in [4]. A synthetic minority oversampling technique is used to increase the class samples. Then, the most reliable features are selected using correlation-based feature selection and fed to the NB classifier for classification.

An efficient approach for gene selection and prediction of cancer is described in [5]. The intelligent feature selection method is the solution for removing irrelevant genes or features from the data. It optimizes the computational cost using different bio-stimulated method called correlation-based feature selection method. The linear SVM and multilayer perceptron classify the microarray data. Relief-F and PSO algorithm based system for DNA classification is implemented in [6]. The Relief-F initially pre-filters the feature and helps to delete the genes with low correlation for the classification of targets, followed by PSO algorithm. Finally, the classification accuracy for the SVM classifier is used as the evaluation function for the feature subsets and obtains the final optimal gene subset.

The chi-square method and SVM-Recursive Feature Elimination method is discussed in [7] for DNA classification. High dimensionality issues are deployed using chi SVM-RFE using the ranking method and the top ten higher weights are considered as important statistical features. The SVM-RFE algorithm selects the informative genes, and Chi SVM-RFE network model is used for the classification. Principal Component Analysis (PCA) and multinomial logit for DNA classification is described in [8]. It uses PCA based extraction of features and a multinomial logit classifier is used for classification of colon cancer, ovarian cancer, pulmonary cancer, and leukemia.

The recurrent neural network-based model using PSO is discussed in [9]. The informative gene-selection method is used for multi-class cancer identification and PSO for selecting the genes relevant to a certain type of cancer. It improves the performance of the classifiers and provides molecular insights for treatment and drug development in cancer-diagnosing scenarios in medical field. Classification Technique as Feature Selection (CTFS) is the new feature selection method discussed in [10] for DNA classification. The extreme gradient boosting and KNN classification techniques aid the CTFS function to select the dominant features by tuning the Bayesian parameter. Three classifiers; NB, Linear SVM, and Random Forest are employed for the classification.

A gene selection method is discussed in [11] for DNA classification by encoding the information of gene-to-class sensitivity. A few discriminative genes are selected by an extreme learning machine. SVM and *k*NN classifiers are used for the classification. A feature selection algorithm is discussed in [12] based on mutual information for DNA classification. To reinforce the feature selection two strategies; feature interaction enhancing and relevance boosting are applied.

An efficient gene selection from microarray data is described in [13] using a hybrid framework by embedded approach and multiple filters. Initially, most relevant genes are selected by an embedded method and use the GA with Tabu search and SVM. The embedded approach is again employed by analyzing occurrence of gene in each subset and further reduces the number of genes. In this study, three different feature selection algorithms; BS, FA and BBS are employed to select the dominant genes by GA and then bagging is employed for DNA classification.

II. METHODS AND MATERIALS

This section discusses a pattern recognition system aimed specifically to predict cancer using microarray data. Pattern recognition can be described as the assignment of a name to an unknown object or event. The name in this context being a label attached to one of a set of classes to which the pattern can belong. The basic problem in these systems is to find suitable methods of forming generalized features about the similarity of patterns belonging to the same class, and also about the difference between patterns belonging to different classes. In this study, GA is utilized to find the best discriminating the sub-set of genes and bagging is employed for the classification. Different stages of a typical classification system are shown in Figure 1.

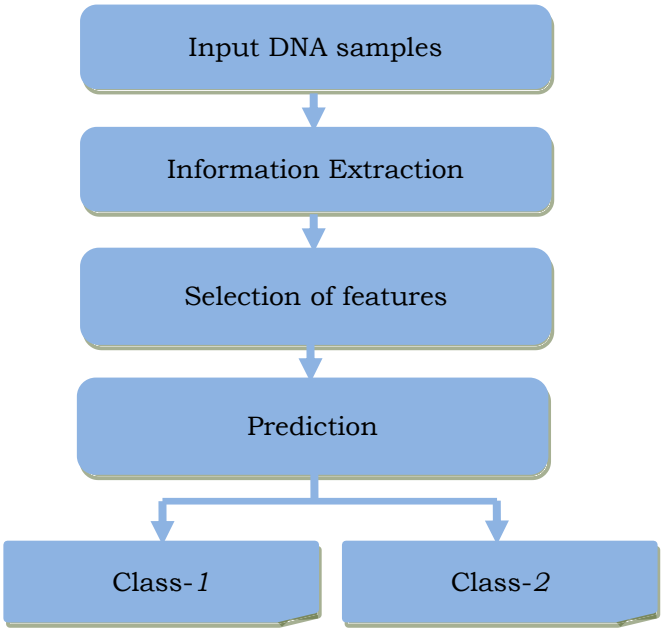

Fig. 1 Different stages of a typical system

A. Genetic Algorithm

The GA is a non-exhaustive adaptive search technique that evolves a population of better individuals based on natural selection. The population in GA is a collection of chromosomes and each chromosome in a population is a potential solution to the problem. The motivation behind a genetic algorithm is the Darwinian evolutionary theory. When the GA is used as a feature selection strategy, the population is initialized randomly with some of the training data used as initial chromosomes. A fitness (evolution step) value is then calculated and only the fit chromosomes (highly discriminative features) are reproduced and the rest are rejected. This process repeats while the fitness between the parent and child population are different which in terms minimizing and maximizing the intra-class and the inter-class distance respectively. The fittest features are maintained and the solution progressively improves through generations. In this context, fitness describes the ability to distinguish between different texture images.

When classification accuracy is used with the GA, feature selection becomes very inefficient since a classifier is trained for every genetically

manipulated feature subset. One feature on a chromosome is represented by one gene and its presence or absence is denoted by a gene with a value of 1 or 0 respectively. The length of a chromosome is the number of features present.

The GA randomly selects a common point in the selected chromosomes and then exchanges their corresponding bits leading to 2 new individuals. This process of exchanging bits is called *crossover.* The main features of a GA are reproduction, and then crossover which is followed by *mutation* where a bit is altered from 1 to 0 or vice-versa based on a specified probability. The mutations and *selection* result in a stepwise optimization of features. Mutation prevents the GA getting stuck in a local optimum. However, mutating the most significant bit can result in a weaker individual. The values of crossover and mutation are typically 0.6 and 0.001 respectively and they are empirically determined and this makes this method suboptimal.

The GAs are recommended for solving complex problems with a large number of features. GAs are relatively sensitive to noise. A GA is an improvement over random and local search methods. Its features are selected as a unit, and the interaction between different features is tested as a group. GA is computationally efficient when used on a larger number of features. Furthermore, it does not need domain knowledge for optimum classification. The application of GA has been in condition monitoring amongst others and the features used were statistical, spectral and wavelets.

A.1 Backward Selection

The BS algorithm selects the best subset of features. Its implementation involves starting with a full set of features. One feature is held back at a time and then all possible combinations of the features in the remaining set are obtained. These subsets are evaluated individually using a criterion function. The best subset is chosen and the process is repeated on this subset to get the next best subset. At each stage the subset which performs best among all other subsets is selected. The number of possible combinations becomes prohibitively high with an increase in the number of features.

A.2 Forward Selection

In contrast to the BS algorithm, the FS algorithm starts with an empty set. It then evaluates individual features and then the best selected feature is added to the empty set. This feature once added is no longer available for evaluation for subsequent selections. The process is repeated until all the best features have been chosen. This technique fails to pick two features that are poor individually, but whose combination gives a highly discriminative performance. This happens when these two features are highly correlated and when the second feature is assumed to be giving little extra information for discrimination. This little information might be crucial for the success in separating the different classes. This technique is less computationally expensive than the BS. There is also an improved version, the sequential forward floating selection (SFFS).

A.3 Branch and Bound Selection

The BBS algorithm generates portions of the solution and computes the criterion for the nodes and in this context, the solutions are subsets of highly discriminative features. Its implementation involves subdividing an initial search region into sub-regions ("branching") which are in turn considered by *bounding* an objective function value and then subdividing in the same way as the initial region.

The goal is to reject large subsets of non-optimal solutions without recourse to exhaustive enumeration to evaluate them. Whenever a suboptimal partial sequence of nodes satisfies a criterion, the sub-tree under the node is rejected and enumeration begins on partial sequences which have not yet been explored.

As an example: let $f(x)$ be the function to be minimized, subject to $x \in X$ and where X is a finite set of possible solutions (set of discriminative features). A list L of outstanding (active) feature subsets is kept and the cost U of the best possible solution found. The recipe for the implementation of this algorithm is as follows:

- Step 1: Set $U = \infty$. Discard any poor solutions. Treat the remaining solutions as one subset. Go to step 3:
- Step 2: Branch step: select one of the remaining subsets and then break it into 2 or more subsets.
- Step 3: bound step: For each new subset, X, compute *l(X)*
- If $l(X) \geq U$, we eliminate X. If $l(X) \leq U$, we reset $U = l(X)$, and X is stored as the best solution so far. The criterion is then re-applied to other subsets until the optimal solution is attained.

In step 1, the best bound rule which partitions the subset with the lowest bound can be used. It aims for an optimal solution and discards larger subsets. Alternatively, the Newest bound rule which partitions the most recently created subset can be used. Its advantage is that it does not jump around the tree too often, hence it is less computationally expensive.

B. DNA Classification

The selection of a proper base classifier becomes a problem as there are many base classifiers available As classifiers using complex features usually performs poor in generalization, simple classifiers should be preferred in designing ensemble classifiers. Decision tree is a simple and unique method in machine learning. Through designing many simple rules step by step, it fits the model of people's understanding of the classification process and it is capable of achieving a highly accurate learning system. Its tree structured decision rule enables informative explanation of the decision rules. Also decision tree can deal with inhomogeneous data which might be describing distinct factors of the object. Thus decision tree is a very general classification approach. After introducing tree pruning schemes, decision tree can be a robust machine learning method with respect to noisy data. All these features make decision tree one of the most popular and fundamental classification systems. The pseudo-code for designing decision trees is as follows:

- Step 1: Choose the feature giving the optimal index;
- Step 2: Divide the training data into groups by using the feature obtained from step 1;
- Step 3: For each group obtained in step 2, repeat the steps 1 to 3 until all the data have been classified.

B.1 Ensemble Method

As compared with standard classifiers, it is the ensemble scheme that makes ensemble classifiers different. The ensemble/fusion method is secondary to the diversity of the base classifiers for successful recognition. Here the bagging method is employed for the DNA classification Bagging is an acronym of bootstrap aggregation. The main idea of bagging is to replicate simple classifiers by voting for the most favored result.

In the bagging method, the training datasets differ slightly by introducing

some re-sampling process to create variations between training datasets for different base classifiers. If the outputs of base classifiers are numerical numbers, they are averaged to generate the output of the ensemble classifier. Otherwise if the outputs are discrete classes, the outputs are selected through majority voting process.. Bagging usually used in many systems because it is simple and reliable. Bagging method thus is still of fundamental importance in designing ensemble classifiers. The pseudo-code of the bagging method is as follows:

Training:
Input: S training set; N - number of bootstrap samples;
$T(\bullet)$ - training algorithm; $f(\bullet)$ - simple classifier.
Main: 1. for i = 1 to N{
2. $S^i \leftarrow$bootstrap sample from S (i.i.d. sample with replacement)
3. $f^i(\bullet) \leftarrow T(S^i)$
4.end }
Output: $f^i(\bullet)$, i= 1, 2,...N
Testing:
Input: x - a new sample
Output: $y^* = \arg\max_{y \in Y} \sum_{i=1}^{N} \delta(y - f^i(x))$

III. RESULTS AND DISCUSSIONS

To analyze the performances of DNA classification system using GA with bagging, publically available cancer microarray data set is used. It consists of colon with 1909 attributes [14] and CNS with 7129 attributes [14], ovarian with 15154 attributes [15], leukemia with 7129 attributes [16], and breast with 24481 attributes [17]. A brief description of microarray datasets is shown in Figure 3.

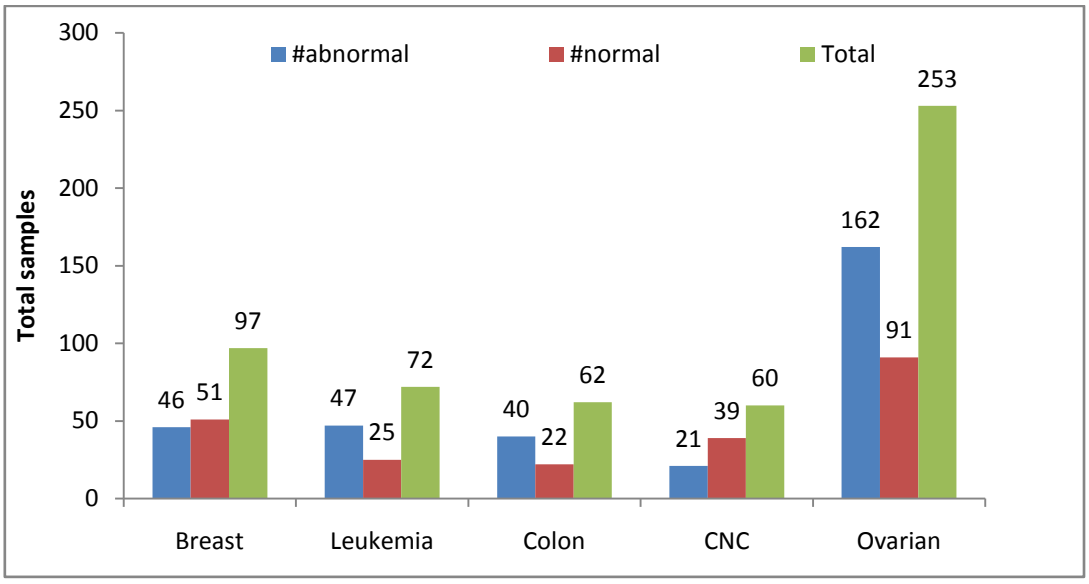

Fig. 3 Descriptions about DNA Database

The performance metrics used in this study are commonly used in medical

domain. They are classification accuracy, specificity and sensitivity. Sensitivity and Specificity refer to the ability of the system to correctly detect the abnormal and normal cases respectively and their formulae are given below:

$$Accuracy = \frac{(T+)+(T-)}{(T+)+(T-)+(F+)+(F-)} \quad (1)$$

$$specificity = \frac{T-}{(T-)+(F+)} \quad (2)$$

$$sensitivity = \frac{T+}{(T+)+(T-)} \quad (3)$$

where T+ is the total number of abnormal cases classified correctly and T- is the total number of normal cases classified correctly. Similarly F- is the total number of abnormal cases classified incorrectly and F+ is the total number of normal cases classified incorrectly. Table 1 shows the performance of GA with bagging algorithm for DNA classification of five cancers; colon, CNS, ovarian, leukemia and breast.

TABLE. 1 Performances of GA with bagging for DNA classification

DNA database	Classifier	Performance measure		
		Accuracy (%)	Specificity (%)	Sensitivity (%)
Colon	GA-BS-Bagging	88.71	86.36	90.00
	GA-FS-Bagging	95.16	95.45	95.00
	GA-BBS-Bagging	100.00	100.00	100.00
CNS	GA-BS-Bagging	88.33	89.74	85.71
	GA-FS-Bagging	91.67	92.31	90.48
	GA-BBS-Bagging	100.00	100.00	100.00
Ovarian	GA-BS-Bagging	92.09	91.21	92.59
	GA-FS-Bagging	96.44	95.60	96.91
	GA-BBS-Bagging	100.00	100.00	100.00
Leukemia	GA-BS-Bagging	86.11	80.00	89.36
	GA-FS-Bagging	93.06	88.00	95.74
	GA-BBS-Bagging	100.00	100.00	100.00
Breast	GA-BS-Bagging	85.57	86.27	84.78
	GA-FS-Bagging	92.78	94.12	91.30
	GA-BBS-Bagging	100.00	100.00	100.00

From Table 1, it is noted that no misclassification occurs for all datasets, the sensitivity and specificity measures maximum using GA-BBS-Bagging. Among the features extracted at different selection approaches, the order of performance is GA-BBS-Bagging > GA-FS-bagging > GA-BS-Bagging. It is also noted that GA-FS-Bagging system provides more than 90% accuracy for all datasets and except Leukemia database, the specificity and sensitivity for all datasets are also more than 90%. Table 2 shows the comparative analysis of the proposed DNA classification system.

TABLE. 2 Comparison of DNA classification system with existing systems

Existing Systems	Accuracy (%)			
	Ovarian	CNC	Colon	Leukemia
[12]	-	-	100	100
[13]	-	94.33	96.75	99.72
[11]	-	-	98.76	100
Proposed system	100	100	100	100

It is clearly observed from Table 2 that the proposed DNA classification system provides promising results than existing systems using different features. Though the system in [11] gives 100% accuracy for Leukemia cancer classification, their accuracies (98.76%) on colon cancer classification are lesser than the proposed system (100%). The proposed system provides ~5% more classification accuracy for CNS cancer compared to one in [13]. It has been known that the quality of the classification process depends greatly on the quality of the features. Results prove that the GA with bagging algorithm performs well for cancer prediction using microarray data.

IV. CONCLUSION

In this paper, an efficient framework for DNA classification is discussed using GA with bagging algorithm. In any medical diagnosis system, the highest priority must be accuracy of the diagnosis, which means that the prediction system with the highest values of sensitivity and specificity must be considered to be the 'best' prediction system regardless of ease of use, time to train or mathematical elegance. Based on the results of this investigation, it is concluded that GA-BBS with bagging is the 'best' DNA classification system. Results show that the system classifies the cancer microarray data with more than 96% of accuracy for all cancer datasets by GA-FS-Bagging system and 100% by GA-BBS-Bagging system.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1]. K. Yan and H. Lu, "An Extended Genetic Algorithm Based Gene Selection Framework for Cancer Diagnosis," 9th International Conference on Information Technology in Medicine and Education, 2018, pp. 43-47.

[2]. P. Wu and D. Wang, "Classification of a DNA Microarray for Diagnosing Cancer Using a Complex Network Based Method," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 16, No. 3, 2019, pp. 801-808.

[3]. I. Jain, V. K. Jain and R. Jain, "An improved Binary Particle Swarm Optimization (iBPSO) for Gene Selection and Cancer Classification using

DNA Microarrays," Conference on Information and Communication Technology, 2018, pp. 1-6.

[4]. N. Soleymani and M. H. Moattar, "An approach based on resampling and feature selection to improve the classification of microarray data," 6th Iranian Joint Congress on Fuzzy and Intelligent Systems, 2018, pp. 61-64.

[5]. B. Patra and S. S. Bisoyi, "CFSES Optimization Feature Selection with Neural Network Classification for Microarray Data Analysis," 2nd International Conference on Data Science and Business Analytics, 2018, pp. 45-50.

[6]. M. Liu, L. Xu, J. Yi and J. Huang, "A Feature Gene Selection Method Based on Relief and PSO," 10th International Conference on Measuring Technology and Mechatronics Automation, 2018, pp. 298-301.

[7]. T. Almutiri and F. Saeed, "Chi Square and Support Vector Machine with Recursive Feature Elimination for Gene Expression Data Classification," First International Conference of Intelligent Computing and Engineering, 2019, pp. 1-6.

[8]. A. Khoirunnisa, Adiwijaya and A. A. Rohmawati, "Implementing Principal Component Analysis and Multinomial Logit for Cancer Detection based on Microarray Data Classification," 7th International Conference on Information and Communication Technology, 2019, pp. 1-6.

[9]. R. Xu, D. Wunsch II and R. Frank, "Inference of Genetic Regulatory Networks with Recurrent Neural Network Models Using Particle Swarm Optimization," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 4, No. 4, 2007, pp. 681-692.

[10]. M. Atlam, H. Torkey, H. Salem and N. El-Fishawy, "A New Feature Selection Method for Enhancing Cancer Diagnosis Based on DNA Microarray," 37th National Radio Science Conference, 2020, pp. 285-295.

[11]. F. Han, C. Yang, Y. Wu, J.S. Zhu, Q.H. Ling, Y. Q. Song and D. Huang, "A gene selection method for microarray data based on binary PSO encoding gene-to-class sensitivity information," IEEE/ACM transactions on Computational Biology and Bioinformatics, Vol. 14, No. 1, 2017, pp. 85-96.

[12]. J. Tang and S. Zhou, "A New Approach for Feature Selection from Microarray Data Based on Mutual Information," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 13, No. 6, 2016, pp. 1004-1015,.

[13]. E. Bonilla-Huerta, A. Hernández-Montiel, R. Morales-Caporal and M. Arjona-López, "Hybrid Framework Using Multiple-Filters and an Embedded Approach for an Efficient Selection and Classification of Microarray Data," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 13, No. 1, 2016, pp. 12-26. 75.

[14]. L. Scott, P. Pomeroy, P. Tamayo and G. Michelle, "Prediction of Central Nervous System Embryonal Tumour Outcome Based on Gene Expression," Letters to Nature, vol. 415, 2009, p. 436-442.

[15]. E. Petricoin, A. Ardekani, B. Hitt, P. Levine, V. Fusaro and S. Steinberg, "Use of proteomic patterns in serum to identify ovarian cancer," Lancet, Vol. 359, No. 9306, 2002, pp. 572–577.

[16]. T. Furey, N. Cristianini, N. Duffy, D. Bednarski, M. Schummer and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using micro-array expression data, Bioinformatics, Vol. 16, No. 10, 2000, pp. 906-914.

[17]. J. Zhang and H.W. Deng,"Gene selection for classification of microarray data based on the Bayes error," BMC Bioinformatics, Vol. 8, No. 1, 2007, pp. 370-379.