# Application of Nature Inspired Soft Computing Techniques for Gene Selection: A Novel Frame Work for Classification of Cancer.

**Rabia Musheer Aziz** ( rabia.aziz2010@gmail.com )
VIT Bhopal University https://orcid.org/0000-0003-2655-7272

**Research Article**

**Keywords:** Artificial Bee Colony (ABC) Cuckoo Search (CS), Genetic Algorithm (GA), Independent Component Analysis (ICA), Naïve Bayes (NB)

**Posted Date:** January 4th, 2022

**DOI:** https://doi.org/10.21203/rs.3.rs-1121838/v1



**Version of Record:** A version of this preprint was published at Soft Computing on April 7th, 2022. See the published version at https://doi.org/10.1007/s00500-022-07032-9.

# Application of Nature Inspired Soft Computing Techniques for Gene Selection: A Novel Frame Work for Classification of Cancer

Rabia Musheer Aziz

Department of SASL (Mathematics),

VIT Bhopal University, Bhopal- Indore Highway, Kothrikalan, Sehore,-466116 (M.P.) INDIA

**ABSTRACT:** *A modified Artificial Bee Colony (ABC) metaheuristics optimization technique is applied for cancer classification, that reduces the classifier's prediction errors and allows for faster convergence by selecting informative genes. Cuckoo search (CS) algorithm was used in the onlooker bee phase (exploitation phase)of ABC to boost performance by maintaining the balance between exploration and exploitation of ABC. Tuned the modified ABC algorithm by using Naïve Bayes (NB) classifiers to improve the further accuracy of the model. Independent Component Analysis (ICA) is used for dimensionality reduction. In the first step, the reduced dataset is optimized by using Modified ABC and after that, in the second step, the optimized dataset is used to train the NB classifier. Extensive experiments were performed for comprehensive comparative analysis of the proposed algorithm with well-known metaheuristic algorithms, namely Genetic Algorithm (GA) when used with the same framework for the classification of six high-dimensional cancer datasets. The comparison results showed that the proposed model with the CS algorithm achieves the highest performance as maximum classification accuracy with less count of selected genes. This shows the effectiveness of the proposed algorithm which is validated using ANOVA for cancer classification.*

Keywords: *Artificial Bee Colony (ABC)* Cuckoo Search (CS); Genetic Algorithm (GA); Independent Component Analysis (ICA); Naïve Bayes (NB).

## 1. INTRODUCTION

The expression levels of the gene in an organism play a discriminant role in clinical studies and the management of several diseases. Microarray technology is a powerful approach for genomic research that creates many analytical challenges for data scientists. Microarray technology uses sophisticated techniques of biomarkers to identify the expressed informative gene. The experimentation with reliable cancer biomarkers plays an important in the field of clinical diagnosis [1]. Classification and analysis of various genetically linked diseases are possible through Microarray technology, which is the most widely used tool in the prognosis of different types of cancer. Accurate prediction of cancer is significant in contributing to effective treatment for the patients. The cancer classification based on gene expression profiles has provided better transparency for the possible treatment strategies. Recently, because of the increase in data generation and storage, various big data applications gain

attention, which also increased interest in applying them to a wide range of biological problems [2, 3]. The identification of genes plays an important role in detecting cancer diseases and has an essential impact for microarray cancer prediction [4, 5]. Therefore dimension reduction followed by the classification acts as the major process for further analysis. However, there are difficulties in detecting cancer from microarray due to the presents of a large amount of gene expression levels in the human body [6, 7]. The input sets of features (genes) are the main factor that influences the quality of the performance of classification algorithms. If the features are relevant to the class labels, the classifier will be able to create a strong relationship between them. However, in most scenarios, the relevancy of features is often unknown and usually, the input data sets have issues such as irrelevancy and redundancy that are not useful during the knowledge discovery process. Thus, this can hinder the process of producing a positive classification. Machine learning techniques for data reduction require knowledge about relevant features and can substantially reduce the size of data for learning algorithms by reducing unnecessary and redundant features. In general, high-dimensional microarray data sets are difficult to interpret and data interpretation is very essential for the treatment of cancer patients. The initial studies of dimensionalities reduction problem, found that the best test error can be attained through a limited number of features (genes) that directly affect the accuracy rates [8, 9]. In a large feature space, it is common to have irrelevant and redundant genes concerning the class labels. Integrality constraints such as irrelevant and redundant features have the capacity to affect the classification performances. Therefore, this research study developed an approach for genes selection to counter all the mentioned drawbacks.

Defining an optimal decision framework for gene selection is an essential but difficult task in the field of machine learning and medical science from microarray data because the characteristic of each data is different. Recently hybrid machine learning techniques gain popularity and by using suitable ccombinations effectively obtain a few relevant and informative genes (features) [8]. Various researches applied varieties of different data mining techniques with different combinations for the problem of identification of significant genes. Motivated by previous researchersnature-inspireded algorithms are more suitable to find optimal set of features from large and complex data of different domains. Techniques comprised of metaheuristic optimization have a broad range from the process of a local search to learning processes [10]. Nature inspired algorithm by conducting them over the search space there by bringing out its best capabilities able to obtain the best of best solutions.

Othman et al., proposed a hybrid multi-objective cuckoo search with evolutionary operators for gene selection. The evolutionary operators used in this study were double mutation and single crossover operators. Proposed approach was tested on seven publicly available, high-dimensional cancer microarray data sets. The experimental results concluded that the proposed work outperformed cuckoo

search and multi-objective cuckoo search algorithms with a smaller number of selected significant genes [11].

Rasmita Dash et al., proposed hybridized harmony search optimization approach for feature selection in high dimensional data classification problem. Proposed technique select optimal minimum number of top ranked genes that provided good classification. The experental results on four well known microarray datasets showed that performance of proposed algorithms was better than other published algorithm for the same problem [12].

Hybrid approach is used to reduce the computational time and to take the benefits of the different dimension reduction method [13, 14]. Hybrid approaches combine different feature selection and extraction method to reduce the dimension of the data. Different researchers applied different combination of algorithm according to the requirement of different data sets.

Hameed et al., compared the performance of three well-known nature-inspired metaheuristic algorithms, namely binary particle swarm optimization (BPSO), genetic algorithm (GA) and cuckoo search algorithm (CS) with tewelve cancer data sets for gene selection and classification. In terms of accuracy, BPSO outscored GA and CS, according to the study. In comparison to GA and BPSO, CS was able to pick fewer attributable genes and was less computationally complex [15].

Some researchers created a classification framework and utilised it to categorise cancer gene expression patterns using various hybrid gene selection algorithms based on various nature-inspired metaheuristic methodologies, with better outcomes than single approaches. The work not only selected very few features but also reduced computational cost by using the collection of new techniques that produced good performance in classification. A comparison result expresses that the proposed hybrid approach have been successfully applied and excels with other existing methods in terms of accuracy. [16-18]. Therefore in this paper hybrid approach based on Nature Inspired Metaheuristics technique is proposed that can produce an optimal feature space with significant genes to improve the classification performance.

Independent component analysis (ICA) method is used for finding underlying components (features) from multidimensional statistical data. The extracted components of ICA are statistically independent, this property distinguishes ICA from other methods [19]. Recently, ICA feature extraction method gain attention as effective genes reduction technique of microarray data for NB classifier [20, 21]. The embedded algorithm of the NB classifier based on conditional independence hypothesis, ICA technique resolved successfully this condition as the ICA extracted components (features) are statistically independent. The major problem of ICA technique is how to obtained best subsets of features from the extracted genes that enhance the classification accuracy of NB classifier. One of the author of [22] used ICA extraction method with stepwise regression for feature selection on five

benchmark microarray datasets and proposed approach demonstrated in improving the classification performance of NB classifier. In [23] Proposed a three level ensemble approach for cancer prediction and classification. For gene selection firstly ensemble Fisher ratio and T-test after that optimize the gene with PSO-dICA method then classified five cancer microarray data sets and found satisfactory results compare to other published results. In [24] selected the feature subset by using ICA that extracted essential information and at the same time separated the noise as extracted features were an independent component. The results of data sets have shown the effectiveness and improvement of the proposed approach. In [25, 26] the author used nature inspired metaheuristic techniques based wrapper methods with ICA and found ICA increase the classification accuracy of different classification algorithm for cancer microarray profile.

Other than above nature inspired algorithm, cuckoo search algorithm is the most popular swarm intelligence algorithm, that is motivated by the egg-laying behavior of cuckoo birds. Recently, CS algorithm gain more popularity in feature selection [27], multi-objectivetive optimization problem [28, 29], data clustering [30], disease detection [31], path planning [32] and soon. The literature have shown that the CS is an effective approach to solve numerous optimization problems of different domain. For different research problems CS algorithm has been widely used, but at the same time for complexity optimization problems CS still need some improvement on exploration face. On the other hand, ABC is widely used approach for finding best number of features for continuous optimization problems of microarray data [33-35]. ABC approach work with the help of three bees for finding best solution (food source) and gives more accurate results comparison to the other swarm based meta-heuristic algorithm. In ABC technique, three types of bees manage the global search and local search procedure for finding best solution. The global search of ABC algorithm for finding a new solution is better in contrast to the other nature-inspired algorithms. Some authors, to maintain the balance between local search and global search procedures, some improvements with different algorithms in the onlooker bee phase (local search) of ABC are proposed that improved the performance of ABC [36].

**A.** The objective of the paper

- To improved the performance of feature extraction technique (independent component analysis) by using hybrid approach.
- Proposed nature inspired hybrid algorithm for solving the existing gene selection process of microarray data based on a soft computing technique.

- Proposed novel framework of classification to increse the perfomance of the NB classifier algorithm by utilizing ICA with improved ABC algorithm.

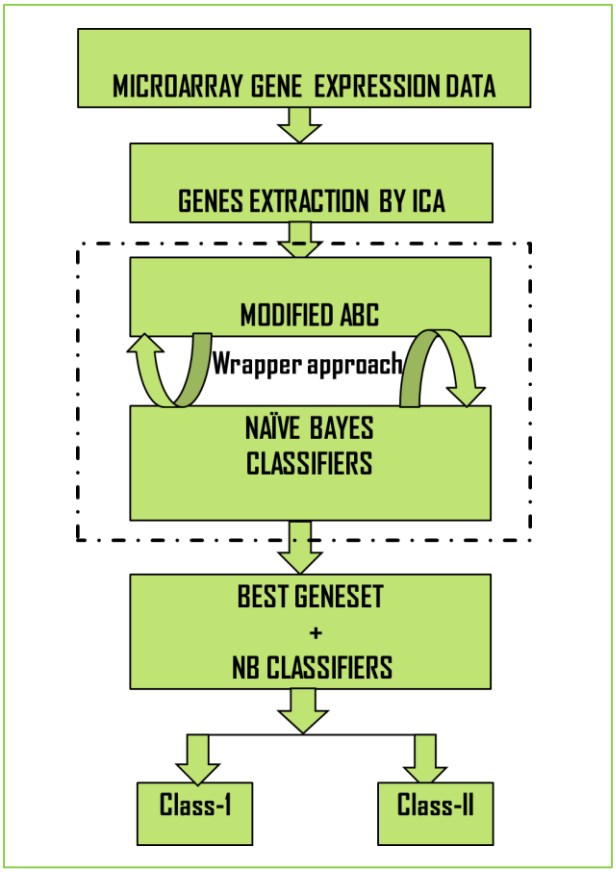


Fig.1. *The proposed framework.*

**B.** Paper Organization

Thus, this article would like to focus on the nature inspired metaheuristic hybrid algorithm in solving the existing gene selection process of microarray data for accurate classification of cancer. The improved ABC algorithm is used to extract the optimal features from ICA feature vectors of microarray data in this research, after that compared the obtained results of proposed framework with others recently published model of NB classifier. The remainder of this research paper is structured as follows, Section two, described the details of feature selection algorithm, Experimental setup is provided in Section three. While Section fourth discussed the experimental results and Section fifth presented the conclusion. Figure 1 shows the proposed framework.

## 2. PROPOSED ALGORITHM

### *2.1* ICA *gene extraction method*

ICA, helps in obtaining hidden features from multi-dimensional information, by decomposing multi variate indications into independent nongaussian sections for the components to be statistically independent [19, 37]. ICA finds correlation between data by decorrelating the data by exploiting or diminishing the distinct in formation. In ICA algorithm all features X treated as a independent components S. If A signify the opposite matrix of a weighted matrix W, and columns of A characterize the source feature vectors of comment X.

$$S = W \times X, \qquad X = A \times S$$

ICA has been extensively utilized for biological information, recognitions and also applicable for other domain. More detailed of ICA extraction method can be found elsewhere [38, 39].

## 2.2 ABC *optimization method*

Recently nature inspired optimization algorithm ABC is popular for genes selection problem of the microarray. The step of ABC approach is based on bees behavior for finding best food source (gene subset). Most of the researchers for the problem of different domain used ABC approach for optimization of the solution. ABC technique work with three classes of bees i.e. employed bees, onlooker bees and scouts bees by using local search and global search techniques [7]. These three classes of bees convergence the problem and find the optimal solution with different effort in the different steps of algorithm. Details information about ABC can be seen in reference [33-35].

## 2.3 Cuckoo search Algorithm(CS)

CS optimization algorithms is based on the reproduction behavior of cuckoo birds. Its developed by using Levy flight rather than the isotropic random walks with infinite variance and mean which cause much longer step from its current position [31]. Basically, the cuckoo lay one or several eggs in other birds nests which aim to ensure the continuity of their generation by directing the host birds to their natural instinct of breed, hatch, and provide food to the baby cuckoos . Three idealized steps of CS are given below [28, 29]:

I. A cuckoo lays one egg at a time and placed its egg randomly chosen nest of other birds.
II. The best nest with the high quality of (solutions) eggs will carry over to the next generations.
III. The availability number of host nests is secured and the egg laid by the cuckoo is detected with the probability, $P_a \in [1, 0]$. While, the host bird either abandon the nest or throw the egg and

form a new nest. The final assumption can be estimated by the fraction $P_a$ of n nests are exchanged with new nests with randomized results. This might increase the survival and reproductive capacity of cuckoo birds, so compare to other algorithm exploitation process of CS algorithm is more efficient [40].

## 2.3 Proposed Algorithm

The searching process of the optimal solution in original ABC approach based on cycle that contains threephases [7].

- Employed bees phase: employed bees search the food sources and estimate their nectar amounts then sending the all information regarding the food sources to the onlookers.
- Onlookers phase: the behavior of onlookers is different of employers, on the basis of received information onlooker make a decision by estimating the nectar amount of all food sources.
- Scouts bee phase: determining the scout because the employed bee of an abandoned food source becomes a scout. Therefore, the employed and onlookers bees manage the exploitation process, on the otherhand scouts bees manage the exploration process in the searchspace.

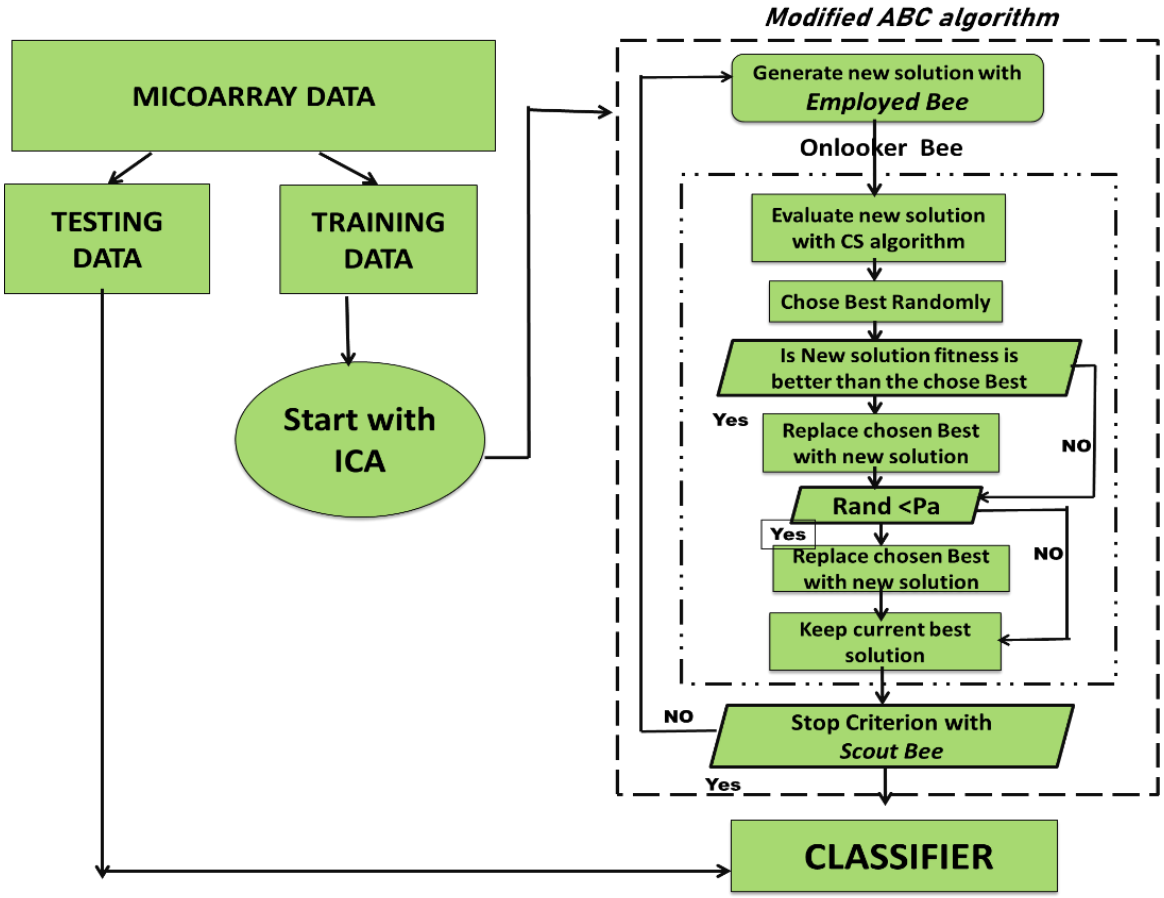


Fig.2. *Illustration of the (CSABC) algorithm works.*

For finding the optimal subset with ABC algorithm, appropriate equilibrium amongst exploration and exploitation is essential [7]. The exploration process of ABC algorithm for finding a new solution compare to the other nature expired algorithm is good, but the exploitation process required more computational time for converge to the optimal solution [25, 41]. Therefore to decrease the computational time of exploitation process of ABC, the proposed algorithm used CS [42]. The CS algorithm has good exploitation process but it ability for searching optimal novel solution in search space is not good compare to ABC algorithm [43, 44]. That is why the CS obtained the local optima too quickly and suffers from the preconvergence problem which is the main issue with CS algorithm [45]. Therefore, to maintain the balanced between exploitation and exploration process, CSABC algorithm uses combination of ABC and CS algorithms. In proposed algorithm, CS algorithm is adopted in the onlooker bee phase as an exploitation process to find the best optimal solution with less computational time by improving the formation sharing between onlooker and employee bees. Furthermore, the idea of using ABC for gene selection with a CS algorithm based on researches [46]. Therefore, the proposed approach uses a combination of nature-inspired metaheuristic algorithms, to reduced disadvantages of the ABC approach such as preconvergence and computational time by maintaining balanced between exploration and exploitation. Figure 2 shows how CSABC works. The code of the CSABC approach is shown in below.

Algorithm for dimensionality reduction

Feature extraction by ICA extraction method

1. Firstly reduced the size of microarray dataset with ICA algorithm.

Optimizition of ICA feature vector with (CSABC) algorithm

Next CSABC algorithm is applied for finding the best genes set from the ICA feature vector for NB classification. A main issue related with ICA is, it normally extracted number of features are equal to the sample size (m), therefore again $2^m$ genes sets exist [47].

Pseudo code of CSABC algorithm:

1. Initialize the population of solutions $x_i \ \forall_i, \ i = 1,2,..., d.$
2. Evaluate the population $x_i \ \forall_i, \ i = 1,2,..., n.$
3. For cycle = 1 to maximum cycle number MCN do
4. Produce and evaluate new solutions from the *employed bees* by using greedy selection process.
5. Calculate the probability values $p_a$ for the solutions $x_i$ by using *CS algorithm in onlooker bees.*

    i. Evaluate its quality/ fitness $x_i$
    ii. Choose a nest among n (say j) randomly if ($x_i > xj$)
    iii. $i^{th}$ the new solution

iv. end if
v. A fraction ($p_a$) of worse nests are abandoned and new are built;
vi. Keep the nests with best quality solutions;
vii. Rank the solution and find the current best;
viii. end while

6. Replace the abandoned solutions with a new one randomly produced xj by using *scout bees*.
7. Memorize the best solution achieved so far.
8. cycle = cycle + 1
9. until cycle = MCN

---

- Return a best solution (important and relevant genes for prediction).
- Train the NB algorithm with best obtained features.
- Classify test set with selected features by using NB classifier.
- Return the classification accuracy.

### 2.4 NB Classifier

NB is famous supervised learning technique in the field of machine learning, it is widely used by many researchers to classify the objects into 2 or more classes by means of using Bayes theorem [48, 49]. It is used widely, when the input variable continuous and independent then the parameters are estimated by the Bayes rule, so that the probability of output variable is exactly predicted. If $E_1$, $E_2$,. . ., $E_n$. are the selected genes from any sample of H, Naïve Bayes classifier classified the samples by using below formula with Bayes theorem as a Naïve Bayes classifier[38, 50]:

$$H' = arg \max_{H \in \omega} P(H) \prod_{i=1}^{n} f(E_i \mid H)$$

Because features of microarray are continuous so for the calculation of class-conditional probability, $f(.|\mathrm{H})$, probability density function with nonparametric kernel density estimation method, for each attributes is used and $P(\mathrm{H})$ is the prior probability of the particular class.

## 3. EXPERIMENTAL SETUP

To evaluate the performance of the proposed approach in this research used six benchmark microarray cancer datasets. In this paper, used six cancer benchmark data sets of gene expression, namely; Colon cancer (Alon et al., 1999), Acute leukemia (Golub et al., 1999), Prostate cancer (Singh et al., 2002),

Lung cancer-II (Gordon et al., 2002), High-grade Glioma data (Nutt et al., 2003) of binary classification and Leukemia 2 (Armstrong et al., 2002) of multi classification. These datasets, downloaded from Kent ridge; an online repository of high-dimensional biomedical datasets (http://datam.i2r.astar.edu.sg/datasets/krbd/index.html). Table 1 shows the full detail and properties of these six datasets.

Table 1 – Detail of six cancer microarray data.

| Data set | No. of classes | No. of genes | Class balance +/- | No. of samples | Short description |
|---|---|---|---|---|---|
| Colon cancer [51], | 2 | 2000 | (22\40) | 62 | Colon cancer data obtained from patients with colon cancer tumor biopsies indicating negative tumors and regular positive biopsies come from healthy areas of the colon of the same patients. |
| Acute leukemia [52], | 2 | 7129 | (47\25) | 72 | Acute Lukemia conation two classes class 1 is the Acute Myeloid Leukemia (AML) with 47 samples and class 2 is Lymphoblastic Leukemia (ALL) with 25. |
| Prostate tumor [53] | 2 | 12600 | (50\52) | 102 | Prostate tumor data collected two class of samples, non-tumor (normal) prostate samples and tumor samples (cancer) . |
| High-grade Glioma [54] | 2 | 12625 | (28\22) | 50 | Glioblastomas and anaplastic oligodendrogliomas of brain tumor samples are contain in High-grade Glioma |
| Lung cancer II [55] | 2 | 12533 | (31\150) | 181 | Tissue samples of Malignant Pleural Mesothelioma (MPM) and Adenocarcinoma (ADCA) of the lung collected in Lung cancer II. |
| Leukemia 2 [56] | 3 | 7129 | (28\24\20) | 72 | Lukemia 2 data set contain three class 28 AML samples, 24 ALL samples, and 20 MLL samples. |

NB classifier uses either kernel density estimation or Gaussian distribution estimation for data classification according to the nature of the data. Since microarray data contain continuous feature, in this paper kernel density approximation is applied with NB classifier [57, 58]. The performance of the proposed approach was examined with two parameters classification accuracy of NB classifier and smallest number of obtained genes and the for all six datatsets. Classification accuracy of the NB classifier is evaluated by the below formula:

$$Classification\ Accuracy = \frac{CC}{N} \times 100$$

Where, CC means correct classified samplesand N is the total number of samples in the respective class. For finiding unbiased results, In this paper implemented Leave One Out Cross Validation (LOOCV) method [59]. To facilitates examination of proposed approach, repeated these experiments with LOOCV at several time. Three other features selection method are also consider for comparing the results of proposed algorithm with the same parameter. The parameters of CS and ABC are adopted with the help of several research paper. For purpose of fair comparison, the same top-ranked genes were chosen for all gene selection method. Experiments were performed for ICA extracted genes in each datasets. To evaluate the performance of the proposed approach in this research all experimental work on preprocessing the datasets has been done using MATLAB 2016b. Moreover, we used python-based tool for feature selection and classification. The experiments were conducted on a desktop computer running 64-bit Windows 10 with an Intel(R) Core (TM) i7-3770 CPU at 3.40GHz and 8GB of RAM. The code of ICA algorithm as the FastICA package, ABC algorithm and CS algorithm are freely available on internet.

### 3.1 Parameter setting and fitness function of proposed algorithm:

Recently some of the researchers used Genetic Bee Colony (GBC) approach for selection of features and found better result compared to original ABC [36]. Genetic Bee Colony (GBC) technique is a combination ABC and Genetic Algorithm (GA) algorithm. GBC is a novel technique, purpose of GBC techniques is to find best subsets of gene for improving the accuracy of different classification algorithm. Therefore, for the sake of a fair comparison, the result of the proposed approach compared with the GBC based hybrid algorithm called (ICA+GBC) with the same classifier. The parameters of proposed approach (ICA+CSABC) and (ICA+GBC) approach that was used in our experiments are given in below tables. For purpose of fair comparison, the same top-ranked genes were chosen for all gene selection method. Experiments were performed for ICA extracted genes in each datasets.

***Parameter setting of proposed algorithm***

| **PARAMETER** | **VALUE** |
|---|---|
| Colony size | 80 |
| Max cycle | 100 |
| Number of runs | 30 runs |
| $Levy\ (s,\lambda)$ | $s^{-\lambda}$ ,1<$\boldsymbol{\lambda}$ <≤ 3 |
| $P_{a_{min}}$ | 0.3 |
| $P_{a_{max}}$ | 0.5 |
| Limit | 5 iterations |

***Parameter setting of (ICA+GBC) algorithm***

| **PARAMETER** | **VALUE** |
|---|---|
| Colony size | 80 |
| Max cycle | 100 |
| Number of runs | 30 runs |
| $\alpha$ *Step size* | 1.0 |
| *Mutation Probability* | 0.02 |
| *Crossover Probability* | o.75 |
| Limit | 5 iterations |

***Evaluating New Solutions and Levy Flight:*** CSABC based feature selection approach find the new optimal solutions with random cuckoo by modifying the parameters using Lévy flight with below equation.

$$x_i^{t+1} = x_i^t + C * Levy\ (s, \lambda)$$

Levy walk generates some new solutions around the obtained best solutions, will accelerate the local search functionality. Here, $C$ has been set to 0.85 from experience. Next evaluate its quality with fitness function of algorithm. **Fitness function** of proposed approach is evaluated by classification accuracy of NB classifier. If the current fitness value is better than the previous one, then skip the previous result, and moves to the current; else it retains the previous solution. Finally, the fitness solution with highest value is returned for best predictive gene subset. The fitness function (fit) is defined as follows:

***Fitness (f) = Accuracy (f)***

**Accuracy (fit):** Testing data **($f$)** classifier accuracy.

***Parameter Pa:*** It represents the probability of discovery of an egg. The *Pa* value is modified dynamically in modified cuckoo search using below equation.

$$P_a = P_{a_{max}} - \frac{P_{a_{max}} - P_{a_{\min}}}{iter_{max}} \times iter$$

$P_{a_{max}}$ and $P_{a_{min}}$ is set 0.5 and 0.3 respectively.

## 4. EXPERIMENTAL RESULTS AND DISCUSSIONS

Table 2 to 7 shows the LOOCV classification accuracy of NB classifiers with different (ICA+CSABC) and (ICA+GBC) approached for above explained six microarray datasets. The best results of all data sets (smallest selected gene size and highest classification accuracy) are highlighted using bold font. Figures 3 (a-d) to 5 (a-d), illustrates the output of the proposed work in the term of AUC curve on six datasets with different threshold, for best subsets genes found by (ICA+CSABC) and (ICA+GBC) approach [60]. From figure 3 (a-d) to 5(a-d) and tables 2-7 gives following results.

Table 2 Classification results of ICA-CSABC and (ICA+GBC) algorithms NB algorithm for Colon dataset.

| No. of genes | Classification accuracy (CA) | | | | | |
|---|---|---|---|---|---|---|
| | (ICA+CSABC) algorithm | | | (ICA+GBC) algorithm | | |
| | **Best** | **Mean** | **Worst** | **Best** | **Mean** | **Worst** |
| 4 | 93.56 | 85.44 | 79.6 | 85.71 | 79.74 | 75.07 |
| 8 | 95.71 | 89.15 | 83.76 | 87.61 | 81.85 | 79.86 |
| 12 | **99.13** | **92.31** | **88.07** | 93.95 | 83.93 | 83.9 |
| 16 | 97.35 | 90.07 | 84.96 | **96.51** | **91.85** | **86.93** |
| 20 | 96.41 | 88.96 | 82.71 | 93.63 | 88.11 | 84.93 |
| 24 | 95.17 | 86.33 | 79.56 | 91.98 | 85.52 | 82.06 |
| 28 | 93.63 | 82.45 | 74.75 | 90.06 | 82.08 | 79.18 |

Regarding colon dataset classification accuracy of test data with the proposed approach achieved 99.13%, 92.31% and 88.07%, best mean and worst respectively which defeated (ICA+GBC ) with best mean and worst by 2.62%, 0.46% and 1.14%, respectively. The best ROC with proposed approach was obtained 96.36 with 12 gene where as ROC value with ICA+GBC was 95.66 with 16 genes for Colon data , which shows that the proposed approach has a discrimination capability between two classes.

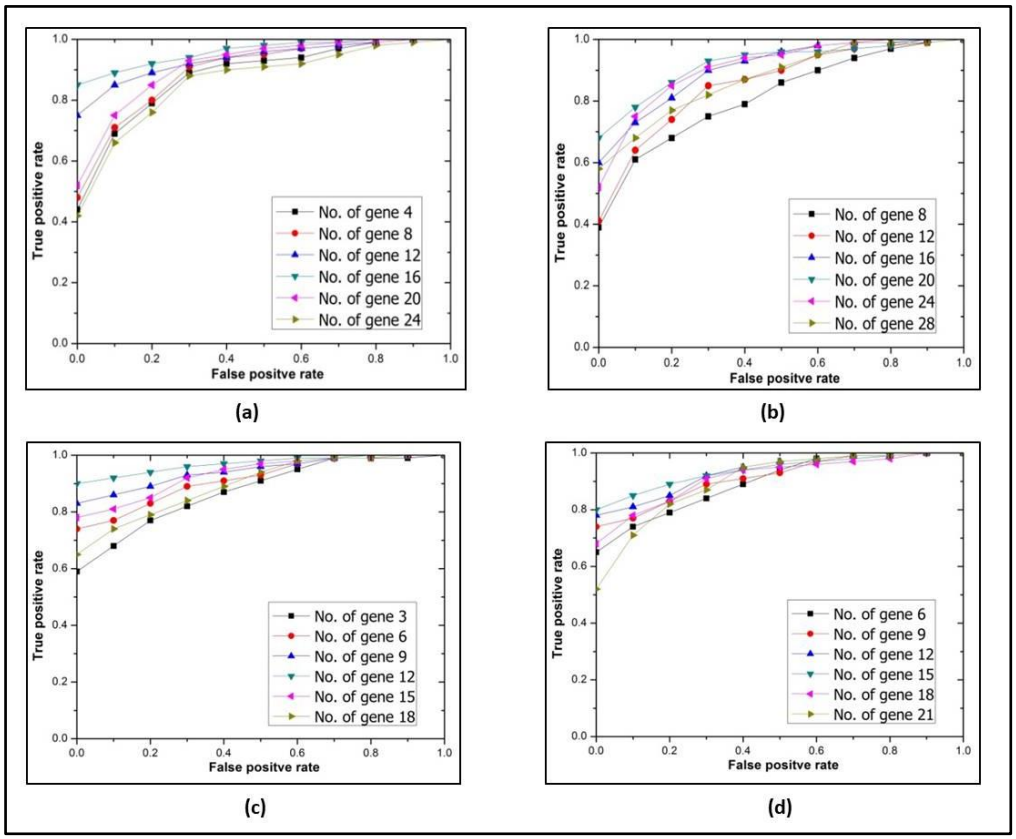

Fig. 3 (a-d *Obtained ROC of (ICA+CSABC) and (ICA+GBC) approach with best number of genes of NB classifier of Colon (a-b) and Acute Lukemia (c-d) dataset.*

Based on the results of test data of Lekumia, the proposed approach obtained 95.24% on an average classification accuracy which is best compared to ICA+GBC and the other hybrid method that used in this experiment. On the other hand the proposed approach obtained 9 important and predictive gene from 72 extracted genes of ICA, which is smallest number of gene compare to obtained genes by the other competitor approaches.

Table 3 Classification results of ICA-CSABC and (ICA+GBC) algorithms NB algorithm Acute Leukemia dataset.

| **No. of genes** | **Classification accuracy (CA)** | | | | | |
|---|---|---|---|---|---|---|
| | **(ICA+CSABC) algorithm** | | | **(ICA+GBC) algorithm** | | |
| | Best | Mean | Worst | Best | Mean | Worst |
| 3 | 90.71 | 81.03 | 77.21 | 89.36 | 80 | 71.51 |
| 6 | 94.01 | 86.06 | 80.02 | 91.26 | 83.54 | 74.21 |
| 9 | **98.97** | **95.24** | **89.14** | 92.52 | 86.16 | 79.88 |
| 12 | 97.27 | 93.84 | 87.34 | **98.41** | **93.11** | **87.65** |
| 15 | 95.05 | 90.75 | 81.53 | 95.51 | 89.51 | 84.45 |
| 18 | 92.66 | 86.74 | 79.15 | 91.63 | 86.5 | 80.05 |
| 21 | 88.45 | 82.71 | 75.07 | 89.8 | 81.56 | 76.3 |
| 24 | 86.77 | 80.35 | 73.22 | 86.22 | 79.47 | 71.21 |
| 27 | 84.71 | 78.83 | 71.38 | 81.42 | 76.65 | 69.11 |

The experimental table 4 and table 8 for prostate data depicted that proposed approach provides 100% best classification accuracy that showed improvement over ICA+GBC, ICA+ABC, ICA+GA, and ICA+PSO respectively for the NB classifier. The AUC value of prostate data found 98.96 with proposed approach for 12 genes.

Table 4 Classification results of ICA-CSABC and (ICA+GBC) algorithms NB algorithm for Prostate tumor dataset.

| **No. of genes** | **Classification accuracy (CA)** | | | | | |
|---|---|---|---|---|---|---|
| | **(ICA+CSABC) algorithm** | | | **(ICA+GBC) algorithm** | | |
| | Best | Mean | Worst | Best | Mean | Worst |
| 4 | 87.26 | 79.93 | 73.97 | 81.55 | 73.2 | 63.65 |
| 8 | 94.62 | 85.85 | 75.88 | 86.34 | 76.64 | 65.75 |
| 12 | **100** | **89.25** | **80.92** | 91.43 | 80.53 | 68.43 |
| 16 | 98.39 | 87.65 | 79.52 | **98.44** | **89.03** | **78.43** |
| 20 | 94.53 | 86.07 | 78.97 | 95.33 | 86.64 | 76.76 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 24 | 92.67 | 82.93 | 75.96 | 93.44 | 83.22 | 71.8 |
| 28 | 91.15 | 81.77 | 74.22 | 91.56 | 81.07 | 69.38 |
| 32 | 90.83 | 80.57 | 72.34 | 89.28 | 79.34 | 68.2 |
| 36 | 88.76 | 78.37 | 71.35 | 86.75 | 77.21 | 66.47 |
| 40 | 86.95 | 76.96 | 69.16 | 85.05 | 75.01 | 63.77 |
| 44 | 85.14 | 74.98 | 67.3 | 82.34 | 72.36 | 61.19 |
| 48 | 84.28 | 73.11 | 65.75 | 81.64 | 71.57 | 60.3 |

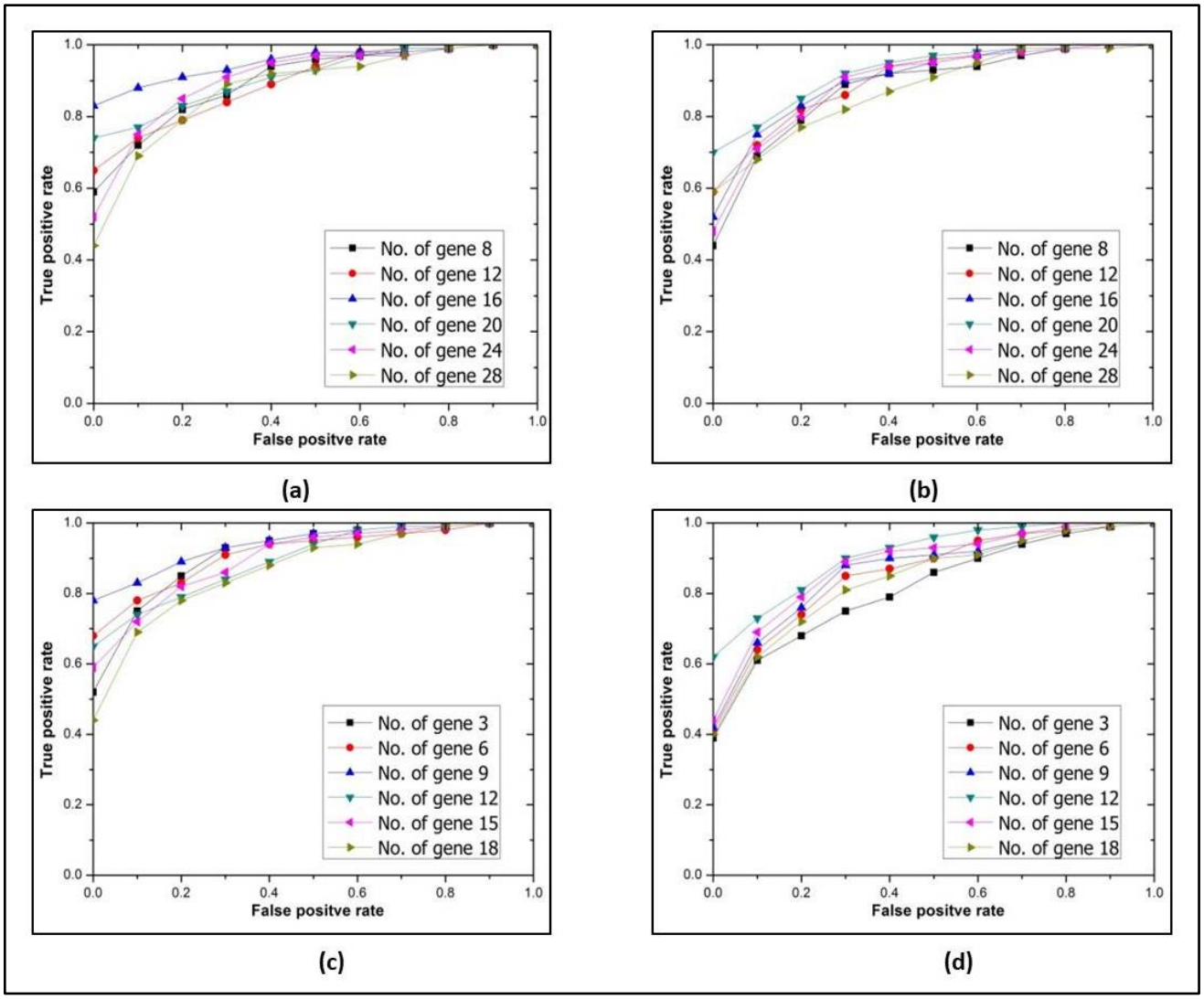


Fig. 4 (a-d). *Obtained ROC of (ICA+CSABC) and (ICA+GBC) approach with best number of genes of NB classifier of Prostate cancer (a-b) and High-grade Glioma (c-d) dataset.*

For the High-grade Glioma dataset, the best training and testing classification accuracy with all ICA features was 87.88% and70.55%, while with the CSABC wrapper method with NB classifier increases training and testing accuracy 98.11% and 96.82% respectively. Therefore, CSABC become a commanding optimization technique for obtaining best feature with ICA for Naïve Bayes classifier.

Table 5 Classification results of ICA-CSABC and (ICA+GBC) algorithms NB algorithm for High-grade Glioma dataset.

| No. of genes | Classification accuracy (CA) | | | | | |
|---|---|---|---|---|---|---|
| | (ICA+CSABC) algorithm | | | (ICA+GBC) algorithm | | |
| | Best | Mean | Worst | Best | Mean | Worst |
| 3 | 91.82 | 85.13 | 77.24 | 88.94 | 80.84 | 71.55 |
| 6 | 93.11 | 86.91 | 79.51 | 92.11 | 85.67 | 78.04 |
| 9 | **97.20** | **90.70** | **83.71** | **96.41** | **89.10** | **80.60** |
| 12 | 94.15 | 88.68 | 82.02 | 90.27 | 85.28 | 79.09 |
| 15 | 90.42 | 85.77 | 79.93 | 88.70 | 82.86 | 75.83 |
| 18 | 88.56 | 83.14 | 76.52 | 86.55 | 80.88 | 74.01 |
| 21 | 87.64 | 81.39 | 73.94 | 85.12 | 79.40 | 72.49 |
| 24 | 86.44 | 79.69 | 71.74 | 82.72 | 78.04 | 72.16 |

With respect to the test dataset of Lung cancer II dataset , proposed approach obtained best classification accuracy 93.45%, that indicates proposed approach applicable to classify all most all samples in their defined classes. Secondly, proposed approach obtained 24 features from 181 extracted features of ICA, which is low compare to other applied approaches.

Table 6 Classification results of ICA-CSABC and (ICA+GBC) algorithms NB algorithm for Lung cancer II dataset.

| No. of genes | Classification accuracy (CA) | | | | | |
|---|---|---|---|---|---|---|
| | (ICA+CSABC) algorithm | | | (ICA+GBC) algorithm | | |
| | Best | Mean | Worst | Best | Mean | Worst |
| 4 | 82.38 | 77.87 | 72.17 | 80.52 | 75.24 | 68.76 |
| 8 | 85.17 | 79.71 | 73.06 | 82.38 | 77.59 | 71.6 |
| 12 | 86.16 | 80.49 | 73.63 | 86.37 | 80.53 | 73.49 |
| 16 | 86.95 | 81.94 | 75.73 | 88.84 | 82.88 | 75.73 |
| 20 | 91.33 | 85.9 | 79.28 | **92.99** | **86.42** | **78.69** |
| 24 | **93.45** | **88.22** | **81.28** | 89.64 | 83.83 | 76.82 |
| 28 | 91.54 | 85.74 | 78.74 | 87.89 | 82.12 | 75.16 |
| 32 | 90.06 | 84.16 | 77.06 | 86.62 | 80.61 | 73.4 |
| 36 | 88.64 | 82.21 | 74.58 | 85.73 | 79.16 | 71.39 |
| 40 | 88.2 | 80.78 | 72.17 | 84.2 | 77.87 | 70.34 |
| 44 | 87.58 | 80.1 | 71.42 | 81.98 | 75.95 | 68.73 |
| 48 | 86.73 | 79.22 | 70.52 | 81.39 | 74.43 | 66.27 |

In the case of multiclass classification of Lukemia 2 dataset, the proposed approach obtained slightly lower classification accuracy compare to ICA+GBC but better classification accuracy compare to other

three method (table 8). But it gives an advantage for obtaining the smallest number of informative and predictive genes, for the NB classifier its obtained 15 genes for the highest classification accuracy which is low compared with 18 obtained genes by the ICA+GBC method.

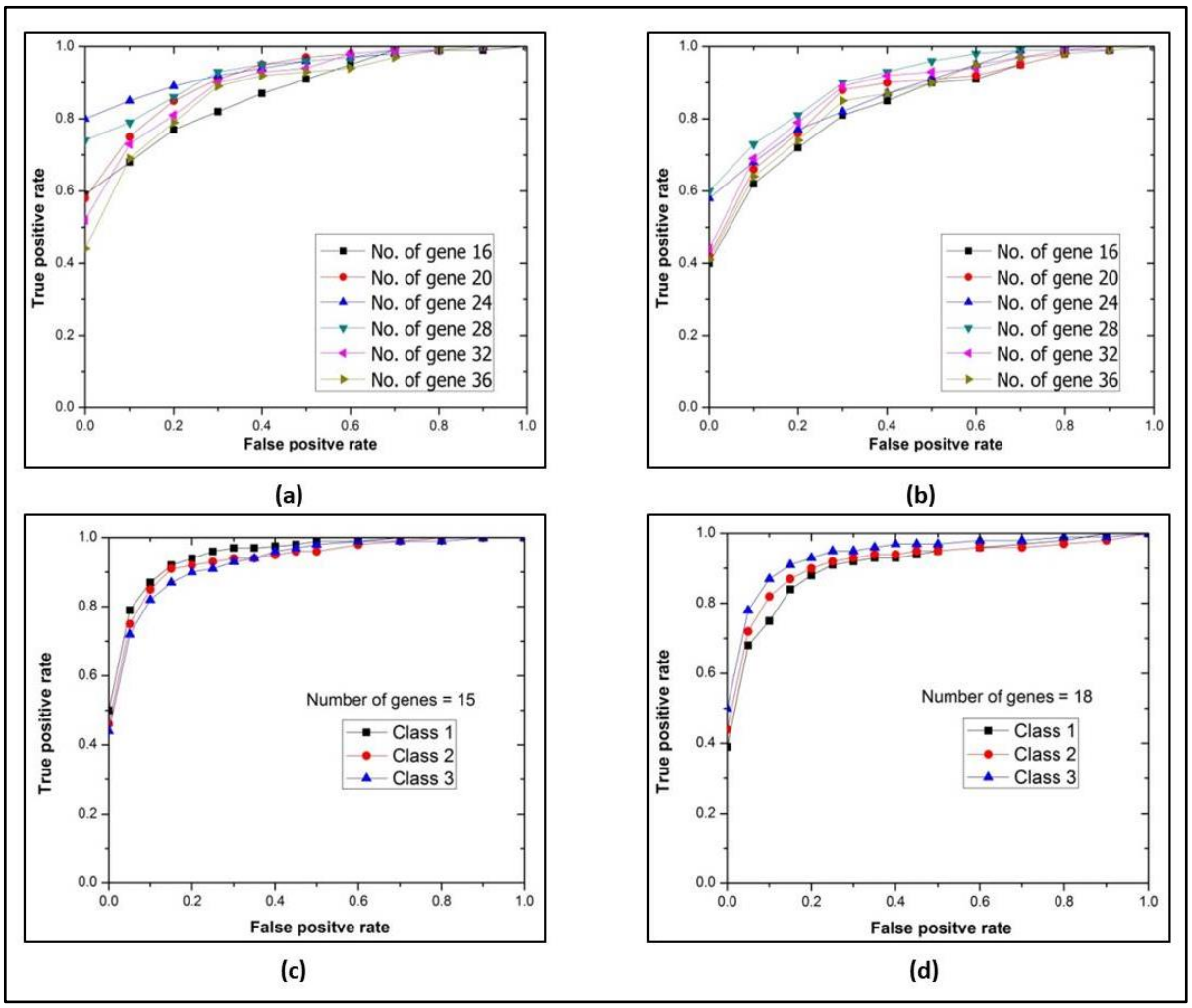


Fig. 5 (a-d). *Obtained ROC of (ICA+CSABC) and (ICA+GBC) approach with best number of genesof NB classifier of Lung cancer II dataset (a-b) and ROC of (ICA+CSABC) and (ICA+GBC) approach with best genes sets of NB classifier of Leukemia 2 dataset (c-d).*

Table 7 Classification results of ICA-CSABC and (ICA+GBC) algorithms NB algorithm for Leukemia 2 dataset.

| **No. of genes** | **Classification accuracy (CA)** | | | | | |
|---|---|---|---|---|---|---|
| | **(ICA+CSABC) algorithm** | | | **(ICA+GBC) algorithm** | | |
| | Best | Mean | Worst | Best | Mean | Worst |
| 3 | 88.38 | 85.47 | 81.36 | 88.33 | 83.17 | 76.48 |
| 6 | 90.07 | 87.93 | 84.6 | 89.05 | 84.64 | 79.03 |
| 9 | 93.92 | 90.24 | 85.37 | 90.05 | 87.61 | 83.98 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 12 | 95.03 | 91.19 | 86.15 | 94.92 | 91.08 | 86.04 |
| 15 | **97.64** | **93.38** | **87.93** | **98.42** | **94.57** | **89.53** |
| 18 | 95.93 | 92.45 | 87.69 | 96.72 | 91.8 | 85.69 |
| 21 | 92.92 | 89.52 | 84.92 | 94.92 | 89.9 | 83.69 |
| 24 | 88.92 | 88.08 | 86.05 | 93.49 | 88.31 | 81.93 |
| 27 | 87.05 | 85.2 | 82.15 | 90.68 | 86.06 | 80.25 |
| 30 | 86.04 | 83.09 | 78.94 | 89.76 | 84.99 | 79.03 |

The proposed technique had effectively performance. Its proved from the experimental results, these outputs were stable and improved compared to the output acquired from the previous experiment. In summary, the proposed technique reached on average 93% for all datasets in terms of classification accuracy which highlight the strength of the proposed technique. Therefore, the experimental result proved, compare to ICA+GBC and other three hybrid methods, the proposed approach has a more significant ability for classifying different samples in their correct classes with NB classifier.

Table 8 displays the ICA+CSABC comparative results with four papular features selection approach over six cancer datasets with the same framework. Experimental results revealed that the ICA+CSABC approach obtained the highest best classification accuracy (97.70 percent) of the six cancer datasets, on the other hand best classification accuracy of the other methods was 96.73 percent, 95.98 percent, 94.42 percent and 92.94 percent for ICA+GBC, ICA+ABC, ICA+GA, and ICA+PSO respectively. This means that the ICA+CSABC algorithm has made the NB classification output more reliable and accurate. The ICA+CSABC algorithm also obtained the smallest number of optimal genes set that contain average (13.50) genes and smallest number of optimal genes set was 14.67, 15.83, 23.67 and 29.17 for ICA+GBC, ICA+ABC, ICA+GA and ICA+PSO. Such findings of the assessment indicate that ICA+CSABC is a promising technique for resolving the problems of dimension reduction and microarray data classification.

Table 8 The classification accuracy of NB algorithm with some nature inspired algorithm when combined with ICA of six cancer microarray datasets.

| **Applied approach** | **Colon cancer** | **Acute leukemia** | **Prostate tumor** | **High-grade Glioma** | **Lung cancer II** | **Leukemia2 data** |
|---|---|---|---|---|---|---|
| ICA+PSO | 91.17(20) | 95.11(19) | 93.31(32) | 91.22(23) | 89.72(41) | 97.22 (40) |

| | | | | | | |
|---|---|---|---|---|---|---|
| ICA+GA | 93.18(18) | 96.58(17) | 95.32(27) | 95.33(18) | 91.84(27) | 94.33 (35) |
| ICA+CS | 92.99(21) | 97.02 (14) | 94.88 (14) | 93.91 (18) | 95.76 (25) | 94.71 (29) |
| ICA+ABC | 98.14(16) | 98.08(12) | 97.08(16) | 94.22(12) | 91.05(24) | 97.33(15) |
| (ICA+GBC) | 96.51(16) | 98.41(12) | 98.44(16) | 96.41(09) | 92.99(20) | 98.42 (18) |
| (ICA+CSABC) | 99.13(12) | 98.97(09) | 100(12) | 97.20(09) | 93.96(24) | 97.64(15) |

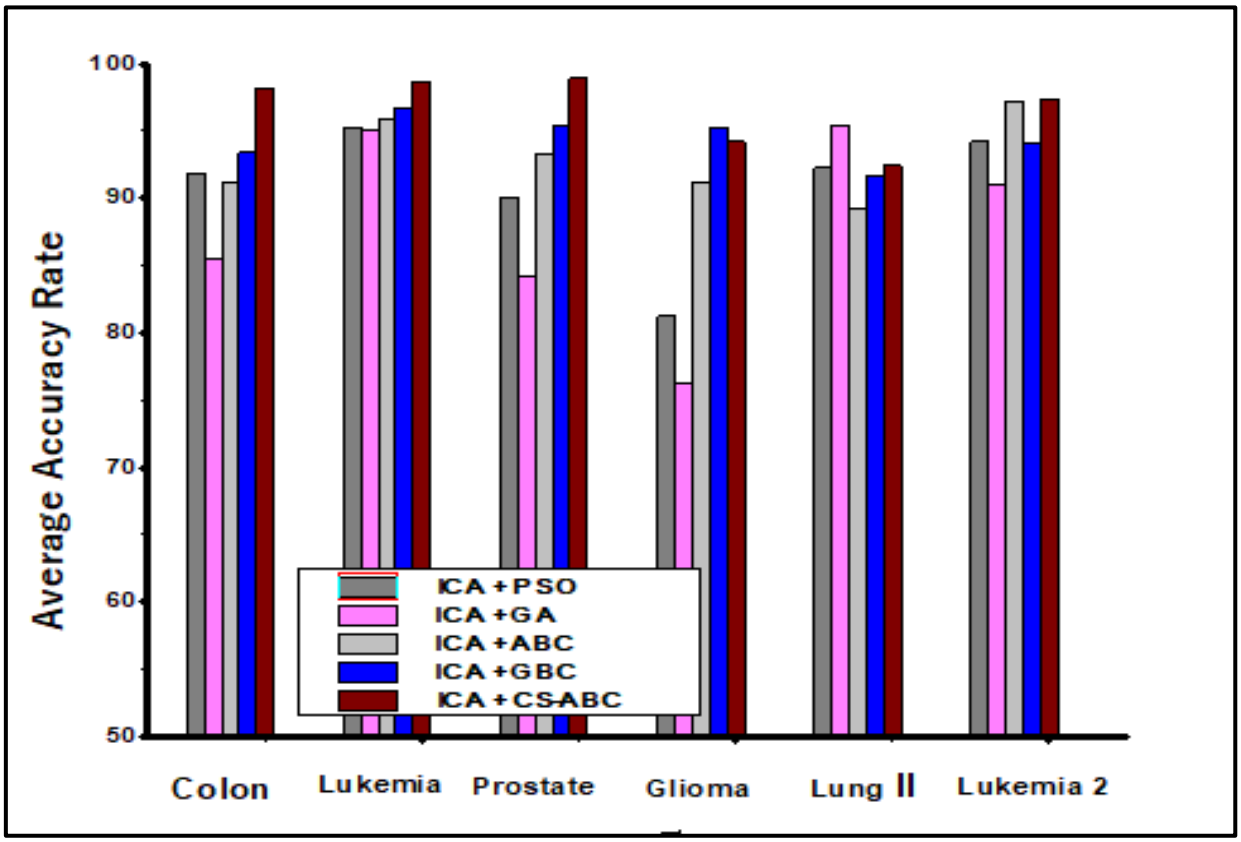


Fig. 6. *Average accuracy rate in six cancer microarray datasets with NB classifier by different features selection approach.*

From the figure 6 the positive outcome with the highest classification accuracy of NB classifier with features obtained by CSABC from ICA feature vectors clearly visible. This is notsurprising because the ICA technique has the ability to resolve the classification criteria of NB classifier.

For The comparision of above five feature selection techniques, a statistical hypothesis test ANOVA was applied with $\alpha = 0.05$ to determine whether there exists a significant difference between them or not. The different approach shown in figure 8 by the name ICA+PSO, ICA+GA, ICA+GBC,

ICA+ABC and ICA+CSABC, are selected as group 1 to 5 respectively in the study. ANOVA tests the hypothesis with $H_0$ and $H_1$.

$H_0$: $\alpha_1 = \alpha_2 = ... = \alpha_5$ (all group means are equal)

$H_1$: not all group means are equal.

In the figure 8 the blue shaded line show the comparison interval for (ICA+CSABC) group mean. Gray line shown the comparison intervals for (ICA+GBC) and (ICA+ABC) and the red line shown the shown comparison interval of (ICA+PSO) and (ICA+GA). The comparison interval (ICA+GBC) and (ICA+ABC) overlap and comparison interval (ICA+GA) and (ICA+PSO) does not overlaps with the comparison interval (ICA+CSABC) group mean. Therefore, the group means (ICA+GBC) and (ICA+ABC) are not significantly different but the group means (ICA+GA) and (ICA+PSO) are significantly different from (ICA+CSABC) group mean.

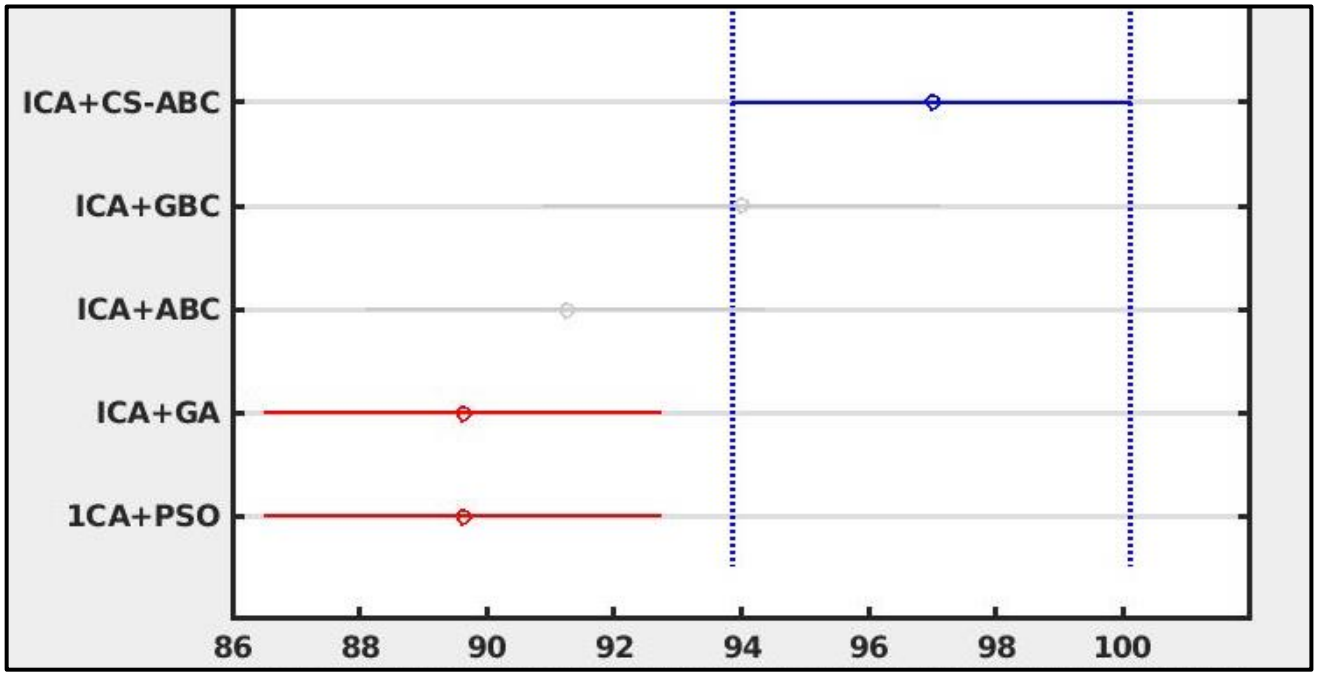


Fig. 8. *The comparision of proposed approach over four different approach with Anova analysis. The red line show the mean significately different from ICA+CSABC algorithm with* $\alpha = 0.05$. *The comparision of*

For further comparisons, the proposed algorithm employed with SVM classifier, because SVM is the best and widely used classifier for microarray data classification as it is less sensitive over the high dimension[61-63]. Since microarray data is a type of non linear classification problem, therefore employed SVM with polynomial kernel, all other parameter of SVM are used on the basis of the research with LOOCV process. Firstly applied ICA technique for feature extraction in the second improved ABC (CSABC) applied to optimize ICA extracted feature with SVM classifier and lastly analyze the classification accuracy.

Table 9 The comparision result of NB and SVM classifier with proposed approach.

| Datasets | NB Classifier | | SVM Classifier | |
|---|---|---|---|---|
| | Mean Classification Accuracy | Number of selected genes | Mean Classification Accuracy | Number of selected genes |
| Colon cancer data | 92.12 | 12 | 92.11 | 11 |
| Acute leukemia data | 93.35 | 09 | 92.45 | 12 |
| Prostate tumor data | 89.14 | 12 | 82.23 | 14 |
| High-grade Glioma data | 90.32 | 09 | 89.22 | 10 |
| Lung cancer II data | 87.71 | 24 | 90.34 | 20 |
| Leukemia 2 data | 93.67 | 15 | 92.33 | 12 |

Table 9 and figure 9 summarizes the classification accuracy and error rate of NB and SVM with proposed approach by using LOOCV iterations for the same parameter settings (CSABC) algorithms. We can easily see from the table 9 SVM with proposed approach aslo produced good classification accuracy which is little bit less or equal to the classification accuracy of NB classifier for Colon, Acute, Prostate, High grate Glioma data and Leukemia 2 datasets. For Colon data set both the classifiers gives same classification accuracy with different number of selected genes. On the other hand for Acute, Prostate and High grate Glioma datasets SVM achived less bit more error rate compare to NB classifier (figure 9). While for Lung cancer II data, SVM achived mean accuracy rate 90.34% 20 genes which is greater then NB classification rate. The error rate of the SVM and NB in classifying Lekumia 2 data were 7.67% with 12 genes and 6.33% with 15 genes, respectively, so in the term of classification accuracy NB classifier is best but the SVM classifier is the best in the term of selected genes for Lekumia 2 data.

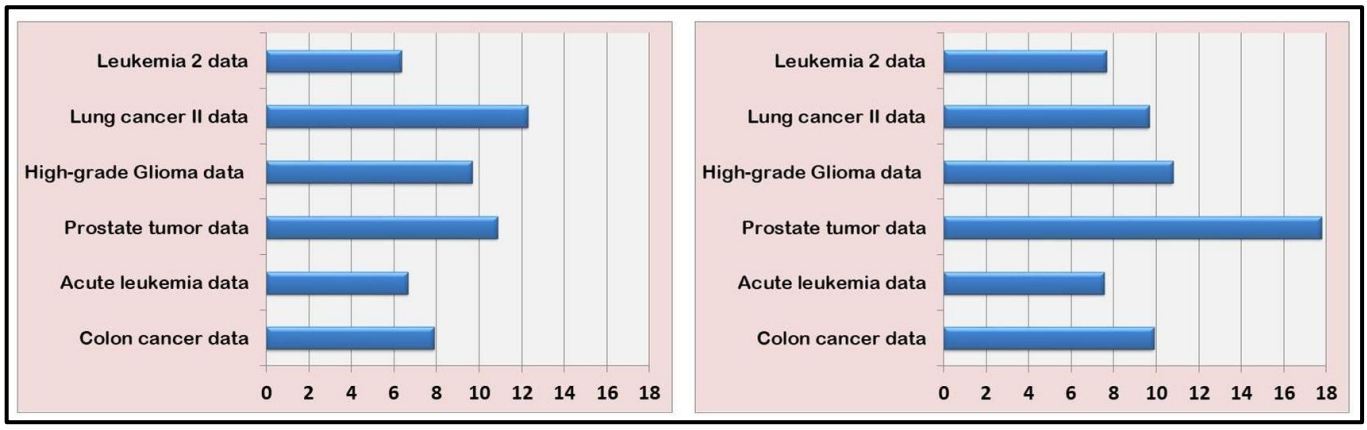


Fig. 9. *Comparision of average error rate of NB and SVM classifiers with proposed feature selection algorithm for all six datasets.*

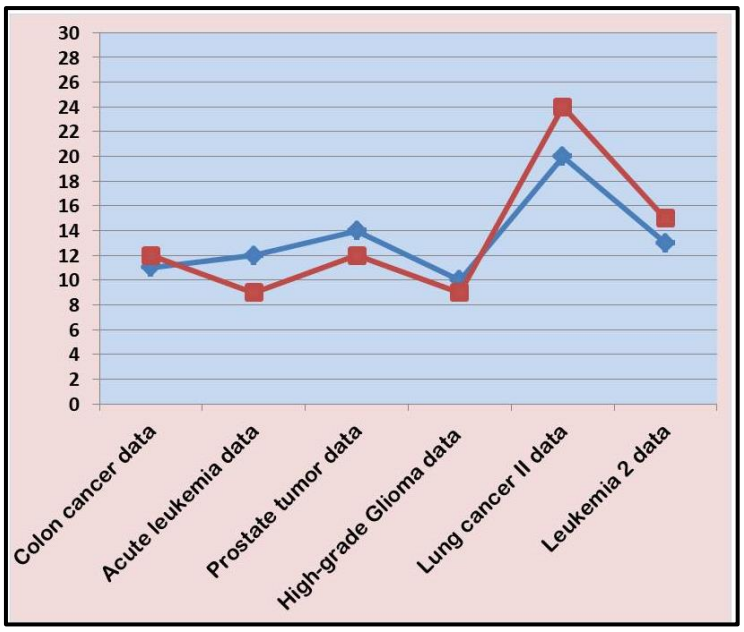


Fig. 10. *Variation of genes for SVM (Blue)and NB(Red) classifier on six datasets with proposed algorithm.*

Figure 10 show smallest number of genes that give best performance of NB classifier, blue and red line show the best number of genes for SVM and NB classifier respectively with all six datasets. The number of obtained genes able to show that the number of important and relevant genes in the cancer microarray data is lesser than 50% of the ICA extracted genes. It can be seen from figure 10, for Colon cancer data out of 62 ICA vectors proposed approach choose only 12 gene for best classification accuracy that show 19.35% of genes were relevant and important for classification. Thus, 80.65% genes are non relevant to disease and cause noise in order to make a good classification model. For other datasets 12.50% of Acute, 11.70% of Prostate, 18.40% of High grade, 13.25% of Lung -II and 20.08% of Lukemia 2 data of ICA feature vector are important for NB classification. Similar result in the term of selected genes found with SVM classifier for all datasets, for Colon and Leukemia 2 data SVM selected 10 and 15 genes which is less compare to NB classifier. For Acute, Prostat and Lung cancer II data SVM selected 12, 14 and 24 genes respectively which is slightly more then NB classifier. Therefore, proposed algorithm identify the important and relevant genes that constructed the best classification model for both the classifiers. The selected number of genes for Lung cancer II data both the classifier required little bit more gene compare to other datasets because every dataset has its own different characteristics.

## 5. CONCLUSION

This paper proposes, novel metahuristic hybreid approach by combining the ICA and advantages of cuckoo search and ABC based feature optimization apprproach with NB classifier for cancer Classification. This method was successes fully reduced the misclassification errors during the classification process on six cancer microarray data. Experimental results show the superiority of the proposed approach in the term of classification accuracy with two factors, best obtained less number of genes set and best AUC score for unbiased accuracy. Therefore, metaheuristic nature-inspired algorithms act as a strong tool in solving microarray cancer data classification problems.

In the future work, incorporating more than one classifier with proposed feature selection techniques to enhance the classification accuracy of the proposed work and to examine the selective classifier mode.

## REFERENCE


1. Ong, H.F., et al., Informative top-k class associative rule for cancer biomarker discovery on microarray data. 2020. 146: p. 113169.
2. Li, G., et al., Prediction of biomarkers of oral squamous cell carcinoma using microarray technology. Scientific reports, 2017. 7: p. 42105.
3. Dwivedi, A.K., Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Computing and Applications, 2018. 29(12): p. 1545-1554.
4. Selaru, F., et al., Global gene expression profiling in Barrett's esophagus and esophageal cancer: a comparative analysis using cDNA microarrays. Oncogene, 2002. 21(3): p. 475-478.
5. Elek, J., K. Park, and R. Narayanan, Microarray-based expression profiling in prostate tumors. In vivo (Athens, Greece), 1999. 14(1): p. 173-182.
6. Salem, H., G. Attiya, and N. El-Fishawy, Classification of human cancer diseases by gene expression profiles. Applied Soft Computing, 2017. 50: p. 124-134.
7. Garro, B.A., K. Rodríguez, and R.A. Vázquez, Classification of DNA microarrays using artificial neural networks and ABC algorithm. Applied Soft Computing, 2016. 38: p. 548-560.
8. Aziz, R., C. Verma, and N. Srivastava, Dimension reduction methods for microarray data: a review. AIMS Bioengineering, 2017. 4(2): p. 179-197.
9. Lv, J., et al., A multi-objective heuristic algorithm for gene expression microarray data classification. Expert Systems with Applications, 2016. 59: p. 13-19.
10. Turgut, S., M. Dağtekin, and T. Ensari. Microarray breast cancer data classification using machine learning methods. in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT). 2018. IEEE.
11. Othman, M.S., S.R. Kumaran, and L.M.J.I.A. Yusuf, Gene Selection Using Hybrid Multi-2020. 8: p. 186348-186361.
12. Dash, R.J.J.o.K.S.U.-C. and I. Sciences, An adaptive harmony search approach for gene selection and classification of high dimensional medical data. Journal of King Saud University-Computer and Information Sciences, 2021. 33(2): p. 195-207.
13. Mafarja, M., et al., Efficient hybrid nature-inspired binary optimizers for feature selection. Cognitive Computation, 2020. 12(1): p. 150-175.
14. Venkatesh, B. and J. Anuradha, A review of feature selection and its methods. Cybernetics and Information Technologies, 2019. 19(1): p. 3-26.

15. Hameed, S.S., et al., A comparative study of nature-inspired metaheuristic algorithms using a three-phase hybrid approach for gene selection and classification in high-dimensional cancer datasets. Soft Computing, 2021: p. 1-19.
16. Baburaj, E.J.I.J.o.S.I.R., Comparative Analysis of Bio-Inspired Optimization Algorithms in Neural Network-Based Data Mining Classification. International Journal of Swarm Intelligence Research (IJSIR),, 2022. 13(1): p. 1-25.
17. Alomari, O.A., et al., Gene selection for microarray data classification based on Gray Wolf Optimizer enhanced with TRIZ-inspired operators. Knowledge-Based Systems, 2021. 223: p. 107034.
18. Kumar, L. and K.K.J.N.C. Bharti, A novel hybrid BPSO–SCA approach for feature selection. Natural Computing, 2021. 20(1): p. 39-61.
19. Hyvarinen, A. and J. Karhunen, Oja., E. Independent component analysis. John Wiley&Sonr, 2001.
20. Hasan, B.M.S., A.M.J.J.o.S.C. Abdulazeez, and D. Mining, A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. Journal of Soft Computing and Data Mining., 2021. 2(1): p. 20-30.
21. Li, J., et al., Multi-source feature extraction of rolling bearing compression measurement signal based on independent component analysis. Measurement, 2021. 172: p. 108908.
22. Fan, L., K.-L. Poh, and P.J.E.S.w.A. Zhou, A sequential feature extraction approach for naïve bayes classification of microarray data. 2009. 36(6): p. 9919-9923.
23. Mollaee, M., M.H.J.B. Moattar, and B. Engineering, A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. 2016. 36(3): p. 521-529.
24. Mahdavi, K., J. Labarta, and J. Gimenez. Unsupervised Feature Selection for Noisy Data. in International Conference on Advanced Data Mining and Applications. 2019. Springer.
25. Aziz, R., et al., Artificial neural network classification of microarray data using new hybrid gene selection method. International Journal of Data Mining and Bioinformatics, 2017. 17(1): p. 42-65.
26. Aziz, R., C. Verma, and N. Srivastava, A Novel Approach for Dimension Reduction of Microarray. Computational Biology and Chemistry, 2017.
27. Pandey, A.C., D.S. Rajpoot, and M. Saraswat, Feature selection method based on hybrid data transformation and binary binomial cuckoo search. Journal of Ambient Intelligence and Humanized Computing, 2020. 11(2): p. 719-738.
28. Cui, Z., et al., A hybrid many-objective cuckoo search algorithm. soft computing, 2019. 23(21): p. 10681-10697.
29. Peng, H., et al., Multi-strategy serial cuckoo search algorithm for global optimization. Knowledge-Based Systems, 2021. 214: p. 106729.
30. Pandey, A.C. and D.S. Rajpoot, Spam review detection using spiral cuckoo search clustering method. Evolutionary Intelligence, 2019. 12(2): p. 147-164.
31. Cristin, R., et al., Deep neural network based Rider-Cuckoo Search Algorithm for plant disease detection. Artificial Intelligence Review, 2020: p. 1-26.
32. Song, P.-C., J.-S. Pan, and S.-C. Chu, A parallel compact cuckoo search algorithm for three-dimensional path planning. Applied Soft Computing, 2020. 94: p. 106443.
33. Musheer, R.A., C.K. Verma, and N. Srivastava, Novel machine learning approach for classification of high-dimensional microarray data. Soft Computing, 2019. 23(24): p. 13409-13421.
34. Coleto-Alcudia, V. and M.A. Vega-Rodríguez, Artificial Bee Colony algorithm based on Dominance (ABCD) for a hybrid gene selection method. Knowledge-Based Systems, 2020. 205: p. 106323.

35. Wang, X.-h., et al., Multi-objective feature selection based on artificial bee colony: An acceleration approach with variable sample size. Applied Soft Computing, 2020. 88: p. 106041.
36. Alshamlan, H.M., G.H. Badr, and Y.A. Alohali, Genetic bee colony (GBC) algorithm: a new gene selection method for microarray cancer classification. Computational biology and chemistry, 2015. 56: p. 49-60.
37. Mollaee, M. and M.H. Moattar, A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification. Biocybernetics and Biomedical Engineering, 2016. 36(3): p. 521-529.
38. Aziz, R., C. Verma, and N. Srivastava, A Fuzzy Based Feature Selection from Independent Component Subspace for Machine learning Classification of Microarray Data. Genomics Data, 2016.
39. Hsu, C.-C., M.-C. Chen, and L.-S. Chen, Integrating independent component analysis and support vector machine for multivariate process monitoring. Computers & Industrial Engineering, 2010. 59(1): p. 145-156.
40. Shehab, M., A.T. Khader, and M.A. Al-Betar, A survey on applications and variants of the cuckoo search algorithm. Applied Soft Computing, 2017. 61: p. 1041-1059.
41. Garro, B.A., K. Rodríguez, and R.A. Vázquez, Classification of DNA microarrays using artificial neural networks and ABC algorithm. Applied Soft Computing, 2015.
42. Zhu, X. and N. Wang, Cuckoo search algorithm with onlooker bee search for modeling PEMFCs using T2FNN. Engineering Applications of Artificial Intelligence, 2019. 85: p. 740-753.
43. Kıran, M.S., et al., A novel hybrid approach based on particle swarm optimization and ant colony algorithm to forecast energy demand of Turkey. Energy conversion and management, 2012. 53(1): p. 75-83.
44. Jatoth, R.K. and A. Rajasekhar. Speed control of pmsm by hybrid genetic artificial bee colony algorithm. in Communication Control and Computing Technologies (ICCCCT), 2010 IEEE International Conference on. 2010. IEEE.
45. Chen, X. and K. Yu, Hybridizing cuckoo search algorithm with biogeography-based optimization for estimating photovoltaic model parameters. Solar Energy, 2019. 180: p. 192-206.
46. Ding, Z., Z. Lu, and J. Liu, Parameters identification of chaotic systems based on artificial bee colony algorithm combined with cuckoo search strategy. Science China Technological Sciences, 2018. 61(3): p. 417-426.
47. Zheng, C.-H., et al., Gene expression data classification using consensus independent component analysis. Genomics, proteomics & bioinformatics, 2008. 6(2): p. 74-82.
48. Friedman, N., D. Geiger, and M. Goldszmidt, Bayesian network classifiers. Machine learning, 1997. 29(2): p. 131-163.
49. Hall, M., A decision tree-based attribute weighting filter for naive Bayes. Knowledge-Based Systems, 2007. 20(2): p. 120-126.
50. Fan, L., K.-L. Poh, and P. Zhou, A sequential feature extraction approach for naïve bayes classification of microarray data. Expert Systems with Applications, 2009. 36(6): p. 9919-9923.
51. Alon, U., et al., Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences, 1999. 96(12): p. 6745-6750.
52. Golub, T.R., et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 1999. 286(5439): p. 531-537.
53. Singh, D., et al., Gene expression correlates of clinical prostate cancer behavior. Cancer cell, 2002. 1(2): p. 203-209.

54. Nutt, C.L., et al., Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer research, 2003. 63(7): p. 1602-1607.
55. Gordon, G.J., et al., Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. Cancer research, 2002. 62(17): p. 4963-4967.
56. Armstrong, S.A., et al., MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nature genetics, 2002. 30(1): p. 41-47.
57. Rabia, A., S. Namita, and K.V. Chandan, A Weighted-SNR Feature Selection from Independent Component Subspace for NB Classification of Microarray Data. International Journal of Advanced Biotechnology and Research, 2015. 6(2): p. 245-255.
58. De Campos, L.M., et al. Bayesian networks classifiers for gene-expression data. in Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on. 2011. IEEE.
59. Xi, M., et al., Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine. Computational and mathematical Methods in Medicine, 2016. 2016.
60. Song, B., et al., ROC operating point selection for classification of imbalanced data with application to computer-aided polyp detection in CT colonography. International journal of computer assisted radiology and surgery, 2014. 9(1): p. 79-89.
61. Belgiu, M. and L. Drăguţ, Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 2016. 114: p. 24-31.
62. Huang, M.-W., et al., SVM and SVM ensembles in breast cancer prediction. PloS one, 2017. 12(1): p. e0161501.
63. Raczko, E. and B. Zagajewski, Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images. European Journal of Remote Sensing, 2017. 50(1): p. 144-154.

## STATEMENTS & DECLARATIONS

**Funding:** The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

**Competing Interests:** The authors have no relevant financial or non-financial interests to disclose. Also the authors declare that they have no competing interests.

**Author Contributions:** Material preparation, data collection analysis and all other work were performed by Dr. Rabia Musheer Aziz

**Data Availability:** All used data are benchmark high dimensional microarray datasets of cancer and are freely available in different repositories.