Transaction on Biomedical Engineering Applications and Healthcare
Vol-1 Issue-1, June 2020, PP.26-33

Hybridizing the Dimensionality Reduction Approaches for Cancer Classification Using Genes Expression Analysis

Ankit Rath[1], Arihant chhajer [2]

[1] B.Tech, Biotechnology, Gandhi Institute of Engineering and Technology, Gunupur, Odisha
[2] B.Tech, Biotechnology, Jaipur National University, Jaipur, Rajasthan
[1]rheaansh@gmail.com , [2]arihantmparcs@gmail.com

***Abstract* - In the DNA microarray datasets, genes expression has made a big impact on the classification of diseases especially in the case of tumor classification. Tumor classification is basically done to predict the cancer on the basis of genes expression profile. Although genes expression dataset are considered to be high dimensional dataset, so dimensionality reduction is very much needed during the classification. In this work to reduce the dimension of genes expression we have proposed the hybrid approach using ReliefF method and the genetic algorithm. The combination of these methods will be used for selecting the subset of the genes before performing the classification. In this work ReliefF method and genetic algorithm will work as a filter method and wrapper method respectively and there combination will form the hybrid method. The results have shown that the proposed work can be implemented on the genes expression dataset to improve the classification accuracy during the disease prediction. The proposed work has computed the classification accuracy of 94.4%, 96.7%, 96.6% and 90.6% on genes expression of Colon cancer, Leukemia, lung and prostate respectively.**

Keywords*— *ReliefF, mutiobjective brain storming, lung cancer, mutation, AUC, accuracy

1. Introduction

Today, machine learning is used very quickly by the health care industry. This mainly helps to detect the use of drugs, prediction of diseases, analyzing the patterns and many more. The main issue in the analysis of health industry datais the high dimensionality because most of the data stored for analyzes in the health sector consist of large numbers of features and dimensions [1]. Gens are one of the main analytical units used in this industry [2]. It consists of a specific nucleotide series, the physical and functional heredity unit.

The chromosome contains 20,000 to 25,000 genes in a human body, which is divided into 46 (2 3 pairs) chromosomes. Sometimes chromosomes can be RNA molecule coded for specific proteins. Gens are expressed as genes, they are formed during DNA transcription and then translated into protein [3]. Gene is expressed through genetic expression. At several stages of gene expression it is not only the primary transcription level but also the post transcription level which regulates the production rate of functional proteins in the cell. Expression of genes is considered a series of sequential steps from transcript to post-translation during protein modifications.

Gene expressions are generally considered to be high dimensional and it is very difficult to classify the high-dimensional data. So data pre-processing is very much needed to classify a high-dimensional dataset. Preprocessing is performed in this form of data with regard to reduction in dimensionality [3]. Reduce dimensionality by choosing features and removing the function. The variable selection, subset variable selection and attribute selection [4] is often referred to as the function selection. This is the way for the most informative knowledge highlights for use in model creation to decrease.

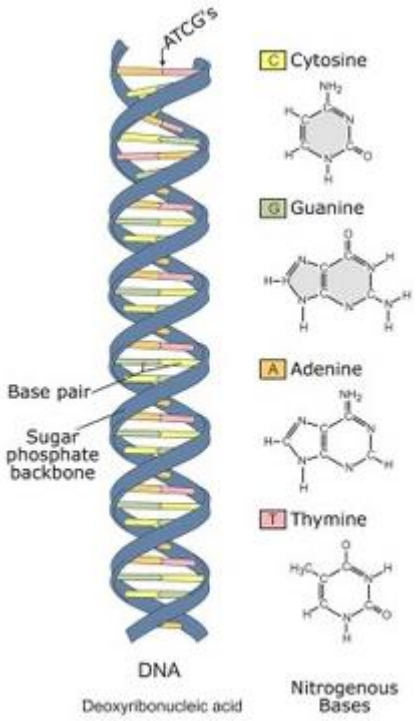

Figure 1. Genes

In genes expression feature selection is regarded as a response to select the most useful highlights that can promote vigorous and accurate machine learning patterns. Informative analysis has a number of procedures. The more modern algorithms are asymptotically superior to previous calculations when measurements are reduced. The feature selection methods are into four approaches which are filter, wrapper, embedded and hybrid approaches [5]. Due to the advantage of high speed and the availability of large datasets, the filter methods have been commonly used, but are easily stuck in local maximum given the classifier's independence.

While the wrapper techniques contain a given learning model, they suffer from high computer expenses, especially in microarray data sets with high dimensions. By using the Markov blanket technique, Wang et al. [6]implemented an enclosure-based gene selection approach to minimize the time required for wrapper evaluation. Filter [7]and wrapper [8] methods does not integrate well with each other and this is considered to be the disadvantage of using the hybrid approaches which may result in lower classification accuracy. This work will show how the hybrid approach can increase the classification accuracy.

In recent times, the Relief algorithm has become one of the high-efficiency filter approach, which has performed well [9]. But the classical relief method is not able to handle the issue of missing data and noise. Kononenko [10] extended Relief as a ReliefF algorithm that can solve multi-classification, missing data and noise and other issues. In this work for performing the feature selection we have used the ReliefF algorithms and it is combine with Genetic algorithms i.e. Multiobjective Brain Storm Optimization Algorithm.

Brain storming algorithm is based on the concept of swarm intelligence and it is inspired by the human activity. Many times it happens that a single person cannot be able to solve the difficult task but that particular task can be solved by the collaborative efforts of the group of individual. Brainstorming algorithm exactly work on this concept in which multiple people works together to solve any difficult problem with good accuracy. It is well known fact that when many people do brainstorming for any difficult problem then their collective efforts can generate the optimal solution for the particular problem.

2. Literature Survey

This section will cover some of the research work done in area of genes expression by using the dimensionality reduction approach.

Hala et al. [11] has used the combination of Genetic Algorithm (GA) and the Artificial Bee Colony (ABC) algorithm. The goal is to combine the advantages of both algorithms. The implemented algorithm is applied to the microarray gene expression profile in order to pick the most accurate and concise genes for cancer classification. Extensive experiments have been carried out to check the accuracy of this algorithm and the research is performed mainly on three binary microarray datasets which includes colon, leukemia, and lung

cancer dataset.

Chandra and Gupta [12] analyzed that to determine the significant characteristics that help to effectively group samples, a careful calculation of the definition of elements is required. In order to select learning characteristics based on relevance and repetition attributes, they presented a new and competent approach to the selection of traits, which depends on the possible amount of measurable characteristics that can be measured for each class called ERGS (Effective Range based gene selection).

Arun Kumar and Ramakrishnan [13] has implemented a customized measure of similarity using a fuzzy rough quick-reducing algorithm for the selection of attributes. The dimensionality reduction in the first stage is done by the Knowledge Entropy based method and the in the second stage they have used fuzzy rough quick-reduction method. Fuzzy rough quick-reduction method defines customizable similar measures to choose the best possible genes in minimum number by removing the unwanted and the redundant genes. The implemented model has used the random forest classifier for performing the classification on the different types of dataset like leukemia, lung and ovarian cancer based on genes expression. The method produces 97.22 percent, 99.45 percent and 99.6 percent accuracy of the leukemia, lung and ovarian cancer gene expression datasets, respectively.

Leclercq et al. [14] developed a framework for comparing knowledge-based and computational selection of genes. In addition, a novel integrative process for the automated combination of the two methods is provided. Results computed from the cancer datasets has shown that the methods based on extrinsic knowledge can compete with complex computational techniques..

Sudipta et al. [15] has used the biological information gained from the Gene Ontology database in the selection of a suitable subset of genes that may further participate in the clustering of samples. The discussed feature selection method was based on unsupervised learning as it does not contain any class label information during the selection process of genes. The clustering approach was implemented to form the cluster with the sample dataset occurred in the reduced gene space.

Cindy et al. [16] Proposed a straightforward approach to combine the external information with conventional gene selection strategies. The developed framework is used for the automatic integration of external knowledge, gene selection and evaluation. The result has shown that the developed framework is a useful tool for evaluation and the ensemble of explicit information will improves the overall results of the analysis.

Songyot Nakariyakul [17] explored the uses interaction information to rank the candidate genes to add in a subset of genes. At a time one gene is added to the current subset and tests whether the resulting subset substantially enhances classification efficiency or not. Only essential genes are selected and the candidate gene list is updated every time a gene is selected, gene was added to the section. Therefore, this gene selection algorithm is very dynamic. Experimental findings on 10 public cancer data sets indicate that the approach is reliable. The research outperforms previous gene selection algorithms in terms of classification precision, despite providing a small number of selected genes.

3. Proposed Work

In this work we have implemented ReliefF algorithm on the genes expression dataset which acts like a Feature estimator that can efficiently handle the complicated issues by computing the relation between the features [18]. The main role of ReliefF method is to select the high quality genes which can helps to perform the classification efficiently by removing the genes which are correlated with each other. In relief method weight is computed in order to efficiently reduce the redundancy in the gene selection and increase the performance of the classification. Let's look at equation given below for computing the distance between the two samples of the same class. Distance between the samples xi on gene subset A of the same class is given by:

$$\text{dist}(A, x_i, H) = \sum_{i=1}^{k} \frac{|xi - \bar{H}|}{\max(A) - \min(A),} \quad (1)$$

In the above equation H represents the distance between the samples and $\bar{H}$ is computed as the average distance among the k nearest neighbors of the samples. max (A) & min(A) represent the maximal and minimal features respectively. Next computing the distance between the sample xi and the sample Mj(C) which is in the different class of the gene subset A and is calculated as:

$$\text{dist}(A, x_i, H) = \sum_{c \neq class(xi)} \frac{p(C)}{1 - p(class(xi))} \sum_{i=1}^{k} \frac{|xi - \overline{Mj(C)}|}{\max(A) - \min(A)} \quad (2)$$

Here p(C) is the division of the target samples C with the total number of samples and the p(class(xi) is the ratio of the samples of classes including xi with the total samples. Mean distance between the k non-nearest samples of the different classes is shown by $\overline{Mj(C)}$.

These above equations are computed to find the Euclidean distance between the K- nearest neighbor points within the same class and between the different classes. If the calculated distance between the points is having the small value that means both the points are closer to each other and the value is greater that means they are far from each other. Euclidean distance computed in this paper is given by the equation:

i. Multiobjective Brain Storm Optimization Algorithm

The other approach that we used in this hybridization approach is the genetic algorithms and the algorithm which is used in this study is the Multiobjective Brain Storm Optimization Algorithm (MBSOA) [19]. Although there are five main steps in Multiobjective Brain Storm Optimization Algorithm (MBSOA) like clustering strategy, generation process, updating global archive, mutation operator and selection operator but we will consider the two main process one is mutations operator and the other one is selection operator, both are used to form the subset of genes. In case of feature selection, MBSOA is based on wrapper methods.

ii. Mutation Operator

Though this process new solutions are derived from the previous ones. Gaussian mutation [20] is applied to form the evolutionary algorithms and the derivation of the new solution can be done by the following process:

$$x_{new}^{d} = x_{selected}^{d} + \xi * N(\mu\ \sigma) \quad (3)$$

$$\xi = \text{logsig}((0.5 * \text{max\_iteration} - \text{current\_iteration})/K) * \text{rand}() \quad (4)$$

In the above equation x_{new}^{d} is the d^{th} dimension of the newly generated subset and $x_{selected}^{d}$ is the dimension of the individual selected genes, Gaussian random function is represented by N (μ σ) with mean μ and σ ; ξ is weight coefficient that contributes the Gaussian mutation; logsig() is a logarithmic sigmoid transfer function, max_iteration is the maximum iteration number and the current_iteration is the current iteration number. K is used for changing the slope of logsig() function and rand() shows the random value between 0 & 1.

There is one more mutation operator other than Gaussian mutation i.e. Cauchy mutation [21], even this can generate better results than Gaussian mutation in some conditions because it has the capability of making longer jumps than Gaussian mutation and this can efficiently use for large class problems. The equation for the Cauchy mutation is as follows:

$$x_{new}^{d} = x_{selected}^{d} + \xi * C(\mu\ \sigma) \quad (5)$$

C (μ σ) is the Cauchy mutation function

iii. Selection Operator

The operator which decides the survival of the newly generated genes for the next generation is known as selection operator. Pareto dominance is used for the selection purpose. There are some rules which has to be followed during the selection of $X_{selected}$ and the mutated individual X_{new} which are as follows:

1. if $X_{selected}$ dominated X_{new} then $X_{selected}$ survives
2. if X_{new} dominated $X_{selected}$ then X_{new} survives
3. if both are not dominating each other then we will randomly select one from the $X_{selected}$ and X_{new} as the new individual.

iv. RMBSOA Algorithm

There are many features in genes expression dataset, so ReliefF Multiobjective Brain Storm Optimization Algorithm (RMBSOA) Algorithm is the new algorithm based on filter-wrapper method that will be used to select the informative genes from the genes expression dataset. This method will help to improve the classification accuracy. Let us look at the flowchart given below.

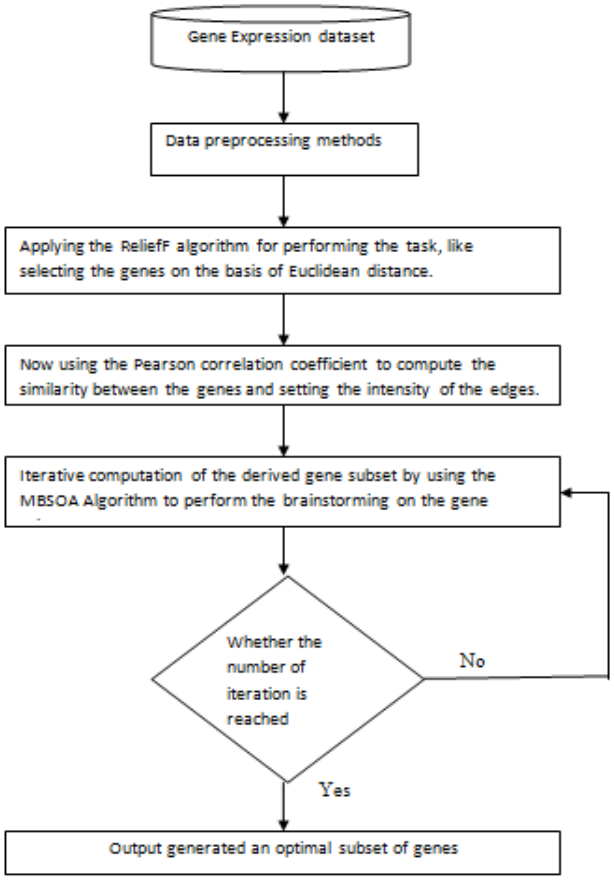

Figure 2: Flowchart of the proposed RMBSOA Algorithm

In the flowchart it has show that how the gene expression with high dimensions is reduced to optimal subset of genes. The preprocessed dataset of gene expression is passed through ReliefF algorithm. The ReliefF algorithm selects the genes on the basis of Euclidean distance and they select the similar genes on the basis of distance. After selecting the genes, Pearson correlation coefficient is used to compute the similarity between the genes and setting the intensity of the genes [22]. Finally the multiobjective brain storming algorithm came into existence and it act as a wrapper selection method. This algorithm will generate the optimal subset of genes on the basis of selecting the genes which are related with the class label only.

4. Results and Discussion

The results are computed by using the Python 3 programming language through Anaconda Framework [23]. All the output are generated using spyder IDE, we have used different datasets of cancer like Colon cancer, Leukemia, Lung and the Prostate cancer for the computation and the results are compared with different types of algorithms like ReliefF, ReliefF+NRS and RFACO-GS on the basis of parameter i.e. classification accuracy.

Datasets	ReliefF	ReliefF+NRS	RFACO-GS	RMBSOA
Colon cancer	78.8%	56.4%	94.0%	94.4%
Leukemia	91.5%	56.3%	95.8%	96.7%
Lung	96.2%	91.9%	99.5%	99.6%
Prostate cancer	93.3%	64.2%	89.2%	90.6%
Average	90.0%	67.2%	94.6%	95.3%

The results are compared on the basis of average accuracies computed on the different types of cancer dataset using the different methods based on ReliefF algorithm. It has been observed that average accuracy generated by using RMBSOA is 95.3% and it is better than the other state of the art algorithms. All the results are shown below using the graph, all the graphs are implemented using Matplotlib library.

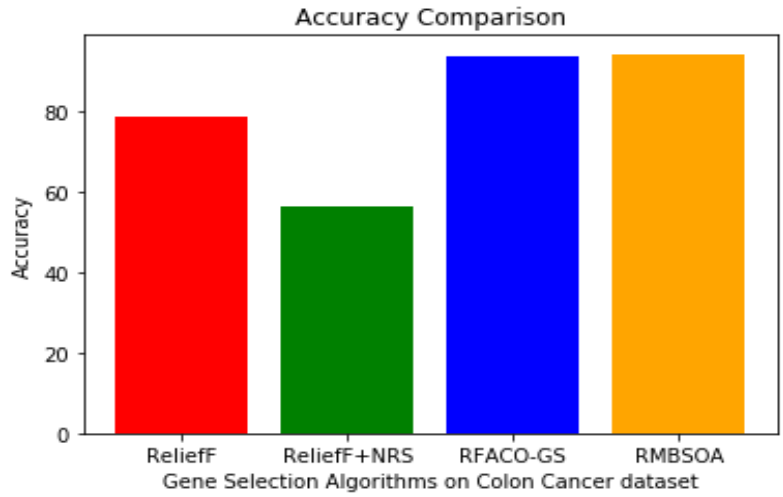

Figure 3. Accuracy comparison of the Colon cancer dataset

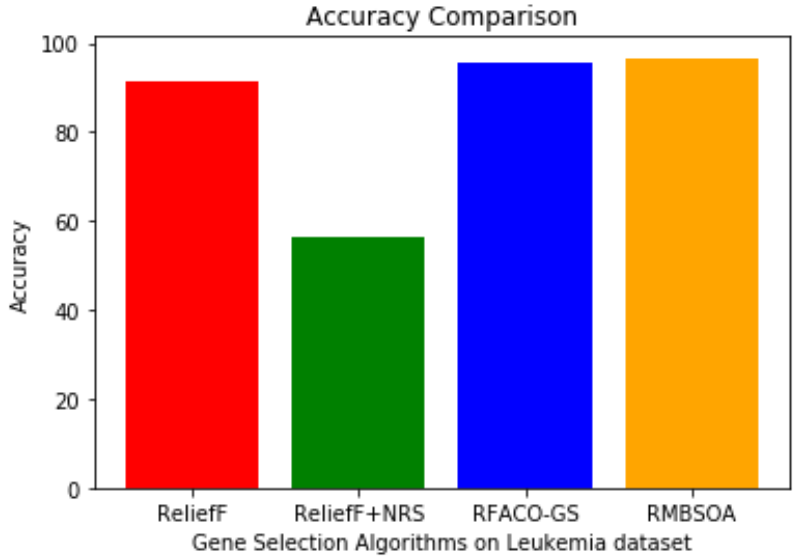

Figure 4. Accuracy comparison of the Leukemia dataset

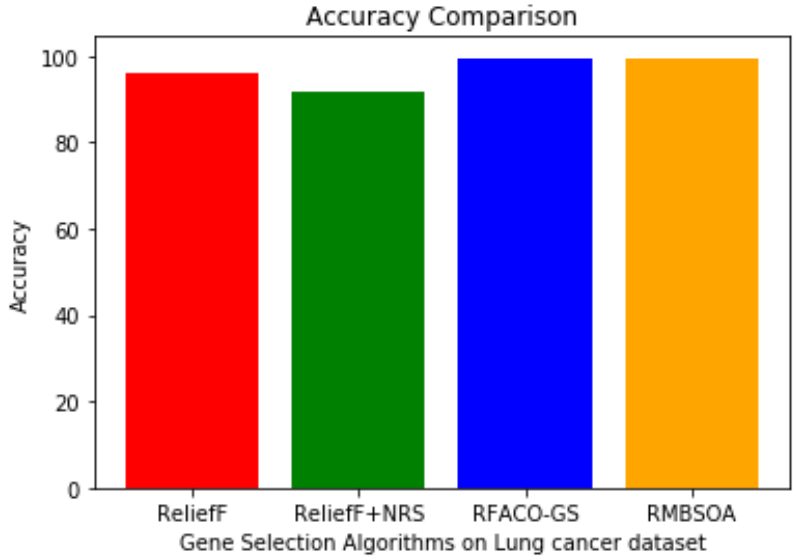

Figure 5. Accuracy comparison of the Lung cancer dataset

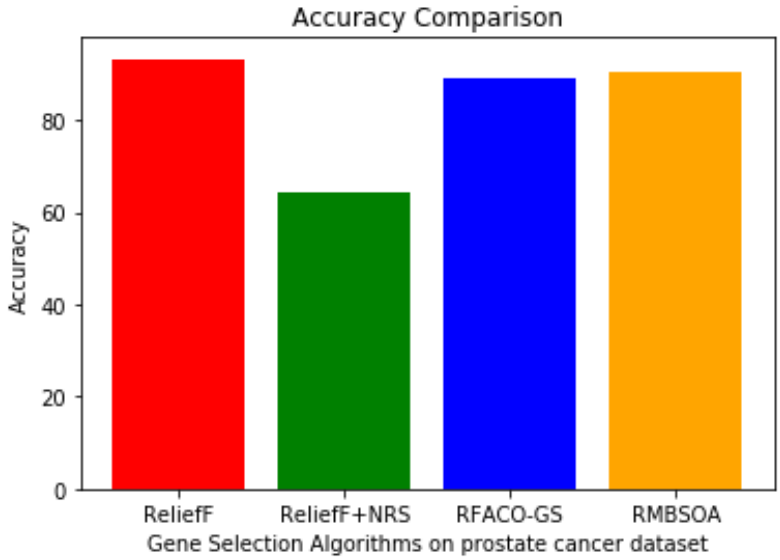

Figure 6. Accuracy comparison of the prostate cancer dataset

Above graphs has shown the accuracy comparison of the models on different types of cancer dataset. The graphs shown in figure 3, figure 4, figure 5 and figure 6 are representing the accuracy of the models on the colon cancer dataset, leukemia dataset, lung cancer dataset and the prostate cancer dataset respectively. From the graphs it can be easily said that that our implemented model i.e. RMBSOA has achieved better accuracy than the other models which are mentioned in the graph.

5. Conclusion

In this work we have implemented the hybrid algorithm using the filter and wrapper method, the filter method is applied by using the ReliefF algorithm and wrapper method is performed by using multiobjective brain storming algorithm. The implemented work has achieved the average classification accuracy of 95.3% which is higher than the other state of the algorithm. Although this work is based on dimensionality reduction but in the future we will use some deep learning approaches because in case of deep learning approaches there is no need of separate dimensionality reduction algorithm. But we cannot guarantee that deep learning algorithm will perform better, it will be totally depend on the dataset.

References

1. M. Verleysen and D. François, “The Curse of Dimensionality in Data Mining,” Analysis, 2005.
2. S. J. Prohaska and P. F. Stadler, “Genes,” Theory Biosci., 2008.
3. R. Nair and A. Bhagat, “A Life Cycle on Processing Large Dataset - LCPL,” Int. J. Comput. Appl., 2018.
4. Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” Bioinformatics. 2007.
5. L. Sun, J. Xu, and Y. Tian, “Feature selection using rough entropy-based uncertainty measures in incomplete decision systems,” Knowledge-Based Syst., 2012.
6. A. Wang, N. An, J. Yang, G. Chen, L. Li, and G. Alterovitz, “Wrapper-based gene selection with Markov blanket,” Comput. Biol. Med., 2017.
7. N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, “Filter methods for feature selection: a comparative study,” IDEAL’07 Proc. 8th Int. Conf. Intell. data Eng. Autom. Learn. Birmingham, UK, vol. 8206, no. December 2007, pp. 178–187, 2007.
8. M. Shardlow, “An Analysis of Feature Selection Techniques,” Univ. Manchester, 2016.
9. R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, “Relief-based feature selection: Introduction and review,” Journal of Biomedical Informatics. 2018.
10. I. Kononenko, “Estimating attributes: Analysis and extensions of RELIEF,” in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 1994.
11. H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, “Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification,” Comput. Biol. Chem., 2015.
12. B. Chandra and M. Gupta, “An efficient statistical feature selection approach for classification of gene expression data,” J. Biomed. Inform., 2011.
13. C. Arunkumar and S. Ramakrishnan, “Attribute selection using fuzzy roughset based customized similarity measure for lung cancer microarray gene expression data,” Futur. Comput. Informatics J., 2018.

14. M. Leclercq et al., “Large-scale automatic feature selection for biomarker discovery in high-dimensional omics data,” Front. Genet., 2019.
15. S. Acharya, S. Saha, and N. Nikhil, “Unsupervised gene selection using biological knowledge: Application in sample clustering,” BMC Bioinformatics, 2017.
16. C. Perscheid, B. Grasnick, and M. Uflacker, “Integrative Gene Selection on Gene Expression Data: Providing Biological Context to Traditional Approaches,” J. Integr. Bioinform., 2018.
17. S. Nakariyakul, “A hybrid gene selection algorithm based on interaction information for microarray-based cancer classification,” PLoS One, 2019.
18. M. Robnik-Šikonja and I. Kononenko, “Theoretical and Empirical Analysis of ReliefF and RReliefF,” Mach. Learn., 2003.
19. S. Cheng, Q. Qin, J. Chen, and Y. Shi, “Brain storm optimization algorithm: a review,” Artif. Intell. Rev., 2016.
20. N. Higashi and H. Iba, “Particle swarm optimization with Gaussian mutation,” in 2003 IEEE Swarm Intelligence Symposium, SIS 2003 - Proceedings, 2003.
21. Q. Wu, “Cauchy mutation for decision-making variable of Gaussian particle swarm optimization applied to parameters selection of SVM,” Expert Syst. Appl., 2011.
22. P. Schober and L. A. Schwarte, “Correlation coefficients: Appropriate use and interpretation,” Anesth. Analg., 2018.
23. wikipedia, “Anaconda (Python distribution),” wikipedia. 2019.