International Journal on Advanced Science Engineering Information Technology

Vol.7 (2017) No. 4-2
ISSN: 2088-5334

An Improved Parallelized mRMR for Gene Subset Selection in Cancer Classification

Rohani Mohammad Kusairi[#], Kohbalan Moorthy[*], Habibollah Haron[+], Mohd Saberi Mohamad[^,$,&], Suhaimi Napis[%], Shahreen Kasim[@]

[#]*Artificial Intelligence and Bioinformatics Research Group, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia*

[*]*Soft Computing & Intelligent System Research Group, Faculty of Computer Systems & Software Engineering, Universiti Malaysia Pahang, 26300, Kuantan, Pahang, Malaysia*

[+]*Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia, 81310, Johor Bahru, Johor, Malaysia*

[^]*Faculty of Creative Technology and Heritage, Universiti Malaysia Kelantan, Karung Berkunci 01, 16300, Bachok, Kelantan, Malaysia*

[$]*Center for Computing and Informatics, Universiti Malaysia Kelantan, City Campus, Pengkalan Chepa, 16100, Kota Bharu, Kelantan, Malaysia*

[&]*Artificial Intelligence and Big Data Institute, Universiti Malaysia Kelantan, City Campus, Pengkalan Chepa, 16100, Kota Bharu, Kelantan, Malaysia*
E-mail: saberi@umk.edu.my

[%]*Department of Cell and Molecular Biology, Faculty of Biotechnology and Biomolecular Sciences, Universiti Putra Malaysia, 43400, Serdang, Selangor, Malaysia*

[@]*Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400, Batu Pahat, Johor, Malaysia*

Abstract— **DNA microarray technology has become a more attractive tool for cancer classification in the scientific and industrial fields. Based on the previous researchers, the conventional approach for cancer classification is primarily based on the morphological appearance of the tumor. The limitations of this approach are the bias in identify the tumors by expert and faced the difficulty in differentiating the cancer subtypes due to most cancers being highly related to the specific biological insight. Thus, this study proposes an improved parallelized Minimum Redundancy Maximum Relevance (mRMR), which is a particularly fast feature selection method for finding a set of both relevant and complementary features. The mRMR can identify genes more relevance to the biological context that leads to richer biological interpretations. The proposed method is expected to achieve accurate classification performance using a small number of predictive genes when tested using two datasets from Cancer Genome Project and compared to previous methods.**

Keywords— **feature selection; cancer classification; mRMR filter method; parallelized mRMR; random forest classifier**

I. INTRODUCTION

In recent technology in molecular biology especially microarray analysis, provide new discoveries in basic biology, pharmacology, and medicine which are facilitated from whole genome RNA. The gene expression of the cell is extremely important to determine its phenotype and function. Thus, the advancement technology in molecular biology had made the cancer classification become a central topic of research in cancer treatment. The classification of a cancer determines appropriate treatment and helps to determine the prognosis [1]. The cancer classification has faced difficulty in part because it relied on specific biological insight rather than systematic and unbiased approaches for recognizing the cancer cells. The approaches for cancer classification are primarily based on the morphological appearance of the tumours [2]. The limitations of the conventional approach are the bias in identify the tumours by expert and faced the

difficulty in differentiating the cancer subtypes due to most cancers being highly related to the specific biological insight. Therefore, microarray technology has been growing recently in order to overcome the limitations in the cancer classifications.

In the recent years, many researchers have observed that microarray gene expression data has an important role in disease diagnosis helping physicians to choose in advance the suitable treatment plan for patients. Recently, biomedical studies had faced a few challenges in order to classify the samples into two states which are disease or not-disease. Due to the challenges, microarray data has become crucial because genetic information in the form of microarray data can be used to identify and classify a new patient's sample into disease or no disease classes. However, considering the amount and complexity of the gene expression data, it is hard to analyse it manually. Thus, there is a need for computational methods. This is a classification problem where the task is to learn a classification model based on a set of labelled training samples from the two populations, and then using learned model to predict the label of test (unseen) samples. Each sample is represented in terms of numeric values obtained from some gene expression measurements. Interpreting gene expression data is not a trivial task because gene expression data is high dimensional in nature, but its sample size is very low. Gene selection is an important sub-problem in such studies. In microarray data, the size of feature-set (gene expression levels) is very much larger than the training sample size [3].

In microarray analysis technology, classification has been a well-known approach that is used to obtain a biological interpretation from the microarray data, and it plays an important role in cancer diagnosis. Basically, in classification, microarray dataset is divided into training data and testing data where the training data is used to train a classifier and this classifier is used to classify the validation data into the correct class of sample. Classification of patient samples (e.g., tumour or healthy) can be used to predict clinical outcome. This is a contrast to the conventional ways of cancer diagnosis that based on examination of stained tissue specimens by a trained pathologist under light microscopy. In recent years, many sophisticated computational methods have been developed based on some well-established classification algorithms such as support vector machine (SVM) by Vapnik [4] and random forest by Breiman [5]. Random forest is a decision tree based ensemble learning method built for both classification and regression while SVM uses a hyperplane with maximized margin to separate the given sample classes. SVM has been widely used in microarray analysis due to its robustness to sparse and noisy data. To that end, a requisite issue in the analysis of this kind of data is to find an appropriate algorithm for selection of the most important features.

Due to the high-dimensional characteristics of the microarray data that with a large number of genes, it is important for the computational methods to select informative genes for more robust classifier in order to achieve better classification performance [6]. This is known as feature selection or more specifically gene selection in the field of bioinformatics. Generally, in gene selection, there consist of three different gene selection approaches namely filter, wrapper and embedded. These approaches differentiate with each other based on their way of integration. Among these approaches, embedded approach integrates the selection method into the classifier, held better interaction with the classifier and better computational complexity [7]. One of the embedded approaches is the hybrid of SVM and smoothly clipped absolute deviation (SCAD) that embedded the SCAD penalty function into SVM that enabled gene selection within SVM for better classification performance [8]. SCAD penalty function was first proposed by Fan and Li [9] that provide a nearly unbiased estimate for large coefficients. Feature selection provides a lot of benefits as it improves prediction performance, understandability, scalability, and generalization capability of the classifier. It also reduces computational complexity and storage, provides faster and more cost-effective model [10], and plays an important role in knowledge discovery. Moreover, it offers new insights for determining the most relevant or informative features [11].

Based on the previous researches, cancer is classified by the type of cell that the tumour cells resemble and is presumed to be the origin of the tumour. Classification of a cancer determines appropriate treatment and helps to determine the prognosis. Thus, the cancer is classified by using mRMR for smaller gene subset selection method by the previous researcher. The lack of the mRMR is the fast and greedy heuristics but it is not guaranteed to find a global optimal solution should one exist. The fundamental problem with redundancy is that the feature set is not a comprehensive representation of the characteristics of the target phenotypes. Moreover, the features selected by a single mRMR run are unlikely to sufficient interpretation for the variety of the biological processes related to the phenotype under study.

In this research, an improved parallelized mRMR for smaller gene subset selection for cancer classification has been proposed. This is because mRMR has fast feature selection method for finding a set of both relevant and complementary features. Among these heuristics, mRMR feature selection technique is particularly appealing because of the relatively low computational complexity of its algorithm for finding the relevant set and the complementary features [12]. An ensemble mRMR implementations outperform the classical mRMR approach in terms of prediction accuracy. Thus mRMR can identify genes more relevance to the biological context that leads to richer biological interpretations. In addition, the parallelized function of mRMR includes in the package of mRMR show significant gains in terms of run-time speed when compared with previously released package. The benefit of MRMR feature set is reduced mutual redundancy within the feature set; these features capture the class characteristics in a wider scope.

The diagnosis of a complicated genetic disease like cancer is normally based on tumour tissue, irrational characteristics, and clinical stages. In treating cancer, early detection can dramatically increase the chances of survival. Thus, time plays a crucial role in treating the disease. Imaging techniques, which are the main method of detection and diagnosis, are only useful once the cancerous growth has become visible. Another common method used to identify

cancer cells is by searching and classifying large amounts of genetic data [13]. There are a few approaches that can be used for feature selection in cancer classification by using Support Vector Machine (SVM), Hybrid GA/SVM, Random Forest and Minimum Redundancy Maximum Relevance (mRMR).

In recent years, DNA Microarray technology can measure the expression level of a great number of genes in tissue samples simultaneously and become promisingly for clinical in the form of diagnosis and prediction of clinical outcomes of cancer and other complex diseases. Therefore, the researcher continuously seeking to develop and finding the most accurate classification algorithms for the creation of a gene expression patient profile in order to maximize the benefits of this technology [14]. Based on Díaz-Uriarte and Andrés [15], the most common task in the gene expression studies is to select the relevant genes for sample classification where the researchers attempt to categorize the smallest gene subset selection that can even now accomplish great prescient performance (for example, for future use with demonstrative purposes in clinical practice).

Cancer is classified by the type of cell that the tumour cells resemble and is presumed to be the origin of the tumour. Classification of a cancer determines appropriate treatment and helps to determine the prognosis. Thus, the cancer is classified by using mRMR for smaller gene subset selection method by the previous researcher. The lack of the mRMR is the fast and greedy heuristics but it is not guaranteed to find a global optimal solution should one exist. The fundamental problem with redundancy is that the feature set is not a comprehensive representation of the characteristics of the target phenotypes [16]-[18]. Moreover, the features selected by a single mRMR run are unlikely to sufficiently interpretation for the variety of the biological processes related with the phenotype under study.

The mRMRe R package which is the minimum redundancy maximum relevance (mRMR) is extended by using the ensemble approach in order to better explore the feature space and build more robust predictors [19]. The ensemble approach had led to computational complexity, thus to overcome the problem, the R package is implemented and parallelized in C using the OpenMP Application Programming Interface. The ensemble mRMR implementations outperform the classical mRMR in terms of the prediction accuracy. In addition, the run-time speed of ensemble mRMR is much better compared to other packages because the parallelized function is included in R package. The mRMRe implements an ensemble variant of mRMR in which multiple feature sets rather than single list features are built. The function for computing a mutual information matrix (MIM) is included in the package based on the appropriate estimators for each variable type (continuous discrete and survival data) [19]. Both this package features have been adapted to fully use multicore platforms. Although mRMR is a fast and greedy heuristic, it is not guaranteed to find a global optimal solution should one exists. To solve these problems, two ensemble approaches had implemented to generate multiple mRMR solutions in parallel; these two techniques are referred to as exhaustive and bootstrap ensemble mRMR.

II. Material and Method

Based on Table 1, the dataset that had been used in this research are obtained by analyzing published pharmacogenomics datasets generated by the Cancer Genome Project (CGP), the Cancer Cell Lines Encyclopedia (CCLE) and CGPS dataset and ALL/AML Leukemia dataset are obtained from the website: datam.i2r.a-star.edu.sg/ datasets/krbd/. The pre-processing of the raw dataset is needed before the data can be used in this research. The extension of the formatted file is .rda file. This format is needed to ensure it can compute in R package environment. Two datasets are used in this research, which are Cancer Genome Project CGP and ALL/AML Leukemia dataset. The features of the dataset shown include a number of genes and number of patients [20]. The CGP dataset consists of 1000 genes with 100 patients while Leukaemia dataset contains 7129 genes of 72 patients.

TABLE I
Microarray Dataset Used

Dataset Name	Genes	Patients	Reference
CGP	1000	100	[17]
Leukaemia	7129	72	[18]

The minimum redundancy maximum relevance (mRMR) with MIQ and MID scheme is described in detail. The genes with significantly different expressions in two different classes (normal and tumor or two different subtypes of cancer) are called differentially expressed genes. The relevance of a gene is referred as the degree of differential expression of that gene. The relevance of gene can be calculated by mutual information [21], [22]. If the expression of a gene has randomly or uniformly distributed in different classes, its mutual information with these classes is zero. If a gene is strongly differentially expressed for different classes, it should have large mutual information. Here we consider mutual information for discrete variable only.

The parallelized functions included in the package show significant gains in terms of run-time speed when compared with previously released packages. The multiple mRMR in parallel is generated. These two techniques are referred to as exhaustive and bootstrap ensemble mRMR [23]. The exhaustive variant enhanced the classical mRMR heuristic by initializing multiple feature selection procedures with the k41 most relevant features. Subsequently, k mRMR solution is produced in parallel, in which the first selected feature is guaranteed to be different. The bootstrap variant resamples (with replacement) the original dataset to generate k bootstraps, and classical mRMR feature selection is performed in parallel for each of the bootstrapped datasets, thus generating k mRMR solutions [24]-[26].

The algorithm that used in this research is maximum relevance and minimum redundancy (mRMR) feature selection algorithm. The conventional of parallelized mRMR method is improved using the maximum relevance and minimum redundancy (mRMR) feature selection algorithm. The algorithm that had been enhanced will implement by using the R package for the best smaller gene subset selection for cancer classification [27].

In this research, two performance measurements are involved to evaluate the effectiveness of proposed method. Firstly, the performance measurement is on the classification accuracy of mRMR method with Random Forest classifier for all microarray datasets. Secondly, the performance measurement is based on run-time performances in seconds of mRMR by using the CGP dataset.

$$q_j = I\left(x_j, y\right) - \frac{1}{|S|}\sum_{x_k \in S} I(x_j\ , x_k) \quad (1)$$

Based on the Equation (1), the set of selected features, denoted by S, is then initialized with xi. Next, another feature is added to S by choosing the feature having the highest relevance with the output variable and the lowest redundancy with the previously selected features, thus maximizing the score q at step j. Step j is repeated until the desired solution length has been attained. This approach had been implemented for continuous/survival, and discrete variables also referred to as F-test Correlation Difference (FCD) and Mutual Information Difference (MID) schemes.

The OOB rate is calculated by selecting the smallest subset of genes with average out of bag (OOB) error rates between the smallest number of genes which is two and the subset with the number of genes that has the lowest OOB error rates. The equation of the OOB rate is shown in Equation (2). In the equation, the p resembles the true efficiency and n is the sample size.

$$\text{Standard error} = \sqrt{p\ (1-p) * \frac{1}{n}} \quad (2)$$

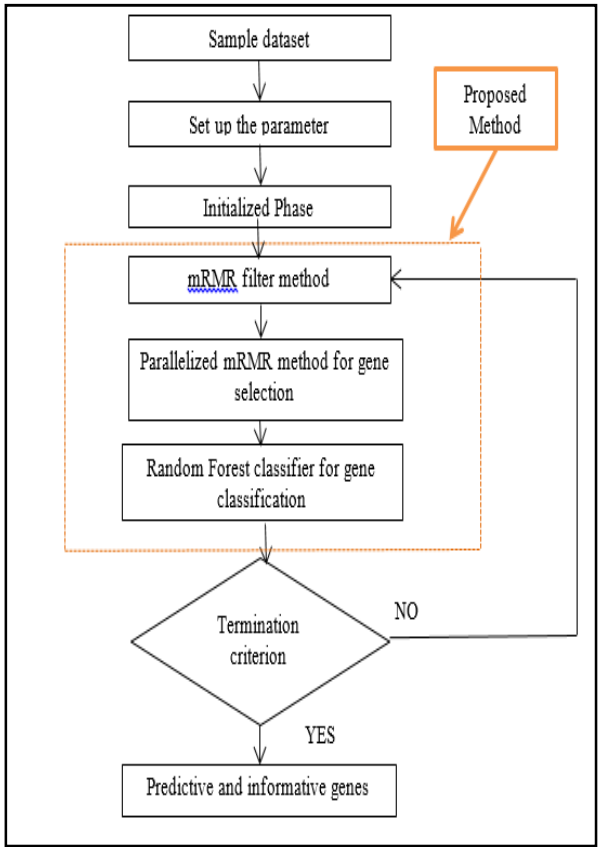

Fig. 1 Proposed parallelized mRMR method

The proposed method is illustrated as in Figure 1. Firstly, the sample dataset that had been used in this research is obtained by analyzing published pharmacogenomics datasets generated by the Cancer Genome Project [CGP; (Garnett *et al*., 2012)] and ALL/AML Leukemia dataset. The dataset is downloaded in txt file format (Leukimia.txt). The dataset is undergoing several steps to convert it into text file format in rda before it can use in R package. All the datasets contain two classes of samples [23]. Then, a few parameter is set up by assigning to a new name and assign certain value during the initializing phase. Next, the initial microarray gene expression profiling is Filtered and pre-processed using the mRMR gene selection method. Each gene is evaluated and sorted using the mRMR mutual information MI operations. The highest relevant genes that give 100% classification accuracy with Random Forest classifier are identified to form a new subset. The mRMR dataset denotes the less redundant and more relative genes as selected by the mRMR approach. The mRMR is applied to filter the irrelevant and noisy genes and reduces the computational load for the mRMR algorithm. The parallelized functions included in the package show significant gains in terms of run-time speed when compared with previously released packages. The multiple mRMR in parallel is generated. These two techniques are referred as exhaustive and bootstrap ensemble mRMR. The exhaustive variant extended the classical mRMR heuristic by initialized multiple feature selection procedures with the k41 most relevant features. Next, k mRMR solutions are produced in parallel, which the first selected feature is guaranteed to be different. The bootstrap variant resamples (with replacement) the original dataset to generate k bootstraps, and classical mRMR feature selection is performed in parallel for each of the bootstrapped datasets, thus generating k mRMR solutions. Next, Use the informative and predictive genes that are generated from the mRMR algorithm in the second phase to train the Random Forest classifier. The Random Forest is applied again to classify the testing microarray dataset and restore the classification accuracy. Lastly, the predictive and informative genes are obtained.

Based on the previous researchers, the mRMR filter method does not involve to remove irrelevant genes and noisy genes. Thus, the number of gene subset selection obtain a large number of genes that lead to computational complexity. In this research, the mRMR filter method is added in order to remove irrelevant genes at the first stage that will reduce the time-consuming. The R package of parallelized is used for gene selection process that improves the run-time speed when compared with the previous packages. Lastly, The Random Forest is applied again to classify the testing microarray dataset and restore the classification accuracy. The improvement had been made by implementing the parallelized mRMR package in R environment and added the Random Forest classifier in order to calculate the accuracy of the gene subset selection.

III. Results and Discussion

Table 2 shows the total number of genes selected using SINGLEGENE, RANK, mRMR and mRMRe. The SINGLEGENE selected only one number of gene but the gene selected sometimes is not the most informative genes,

and it's hard to analyze. Next, the total number of genes selected by RANK and mRMR is 16 but the value of correlation in mRMR greater than RANK. Then, mRMRe selected 214 the number of genes. This selected genes is large and will increase the computational time. The gene is selected based on the correlation value for each gene. Lastly, the improved method had selected only 10 numbers of genes. The gene numbers selected is smaller compared to RANK, mRMR, and mRMRe. Thus, the computational time is much smaller than others. The IC50 is used to measure the sensitivity to detect the cancer cells. This measured sensitivity (IC50) detected Irinotecan (Camptothecin), a therapeutic drug mainly used in colon cancer. This metric is applied to discriminate between resistant and sensitive cell lines.

TABLE II
GENE SUBSET SELECTION [12]

Methods	**Gene Selection**
SINGLEGENE	Total number of genes selection =1
RANK	Total number of genes selection =16
mRMR	Total number of genes selection =16
mRMRe	Total number of genes selection =214
Improved method	Total number of genes selection =10

In this experimental result, the classification has been carried out based on the smallest subset of genes selected in the gene selection process as illustrated in Table 3. The selection is based on the smallest subset of genes with lowest out of bag (OOB) error rates. Based on Table 3, the OOB rate for CGPS dataset is 0.06 and 0.02 for leukaemia dataset. A smaller value indicates higher prediction accuracy, so lower values are better. In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run, as follows:

Each tree is constructed using a different boot-strap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the kth tree. Put each case left out in the construction of the kth tree down the kth tree to get a classification. In this way, a test set classification is obtained for each case in about one-third of the trees.

TABLE III
OOB RATE

Dataset Name	**OOB Rate**
CGP	0.06
Leukaemia	0.02

Classic function of mRMR allows an efficient selection of relevant and non-redundant features. The method ranks X by maximizing the MI with y (maximum relevance) and minimizing the average MI with all the previously selected variables (minimum redundancy). The runtime values below indicate the performance of mRMR. Table 4 shows the runtime using the system in seconds. The lower value indicates the best value that leads to lower computational time. The lower value of runtime because of the smaller gene subset selection. The computational time of the existing mRMR method by using CGPS dataset is 28s while the improved method only takes 2s to execute the system. The leukaemia dataset is added to test on the computational run time. The time taken to execute the system is 7s.

TABLE IV
RUN TIME(S)

Dataset	**Existing mRMR**	**Improved mRMR**
CGP	28	2
Leukaemia	-	7

The experiment result and analysis has been discussed, where the detailed result has been provided according to based on selection method. The result and discussion of each enhanced technique used to achieve the target result have been populated in order to display the differences as well as the improvement achieved. The enhancement method which is a selection of the smallest subset of genes with the lowest OOB error rates.

IV. CONCLUSION

The improved parallelized mRMR method helps to overcome the limitations of the original single mRMR method. Thus, it helps in faster convergence in the conventional single mRMR and contributes in shorten the computational time as well as increase the accuracy of the smaller gene subset selection in cancer classification. In addition, the maximum relevance and minimum redundancy (mRMR) feature selection algorithm also contributes to reducing the noise data. The parallelized mRMR method helps in shortens the computation time compared to the conventional single method. Therefore, the maximum relevance and minimum redundancy (mRMR) feature selection algorithm and parallelized mRMR method increase the accuracy if the smaller subset of genes selection for cancer classification in this research [28], [29].

The difficulties in doing this research are the dataset is complicated since the dataset required to undergoes pre-processing data before it is used in this research. Thus, for the future works, the dataset should be provided in several types of file format or reduced the complexity of the dataset for reducing the time-consuming in pre-processing data. On the other hand, the dataset that contains high-dimensional or complexity can be used to experiment to obtain the accuracy result of the smaller gene subset selection for the cancer classification. In addition, the proposed method can be used for the smaller gene subset selection in cancer classification for the large dataset. Moreover, a comparison of current hybrid methods and the proposed method for observing the differences performance and accuracy of the methods. The t-test can measure the performance of the proposed methods in the future. For the future work, try to reduce the inter-subset distance of the features, expecting to find the best compact gene array which has the most effect on the target classes due to classification task which is a supreme lead in the diagnosis of tragic diseases, e.g., cancers [30]. The further research of mRMR can be a focus on datasets with multiple-class labels. Other statistical measurements such as

information gain, the chi-square test can be considered for gene ranking. Hybrid approaches of optimization may be implemented with an improved solution which can be suggested to avoid premature convergence [30].

ACKNOWLEDGMENT

This research is supported by Universiti Teknologi Malaysia through the Tier 1 research grants (Grant numbers: Q.J130000.2528.11H11).

REFERENCES

[1] G. Bontempi and B. Haibe-Kains, "Feature selection methods for data mining bioinformatics data," Bruxelles, Belgium: ULB Machine Learning Group. 2008.

[2] H. Alshamlan, G. Badr, and Y. Alohali, "mRMR-ABC: A Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling," BioMed research international, 2015.

[3] Babu, M., & Sarkar, K "A comparative study of gene selection methods for cancer classification using microarray data," In Research in Computational Intelligence and Communication Networks (ICRCICN), pp.204-211, Sept. 2016

[4] V. Vapnik, "The Nature of Statistical Learning Theory", New York: Springer, 1995.

[5] L. Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001.

[6] S. Ma and J. Huang, "Penalized feature selection and classification in bioinformatics". Briefings in Bioinformatics, 9(5), 392-403, 2008.

[7] Y. Saeys, I. Inza, P. Larranaga, "A review of feature selection techniques in bioinformatics", Bioinformatics, 23(19), 2507-2517, 2007.

[8] H. H. Zhang, J. Ahn, X. D. Lin, C. Park, "Gene selection using support vector machines with non-convex penalty", Bioinformatics, 22(1), 88-95, 2006.

[9] J. Fan, R. Li, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties", Journal of the American Statistical Association, 96, 1348-1360, 2001.

[10] M. Gutkin, R. Shamir, and G. Dror, "SlimPLS: A method for feature selection in gene expression-based disease classification," PLoS ONE, vol. 4, no. 7, p. e6416, Jul. 2009.

[11] Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A., "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection," IEEE/ACM transactions on computational biology and bioinformatics, 13(5), 971-989, Sept 2016.

[12] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(8), 1226-1238, 2005.

[13] Al-Rajab, M., Lu, J., & Xu, Q., "Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis," Computer Methods and Programs in Biomedicine, 146, 11-24. July 2017.

[14] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification", BMC bioinformatics, 9(1), 319, 2008.

[15] P. K. Ammu, and V. Preeja, "Review on Feature Selection Techniques of DNA Microarray Data", International Journal of Computer Applications (0975–8887) vol, 39-44, 2013.

[16] K. Das, J. Ray, and D. Mishra, "Gene Selection Using Information Theory and Statistical Approach", Indian Journal of Science and Technology,8(8), 695-701, 2015.

[17] C. Ding, and H. Peng, "Minimum redundancy feature selection from microarray gene expression data", Journal of bioinformatics and computational biology, 3(02), 185-205, 2005.

[18] R. Díaz-Uriarte, and S. A. De Andres, "Gene selection and classification of microarray data using random forest", BMC bioinformatics, 7(1), 3, 2006.

[19] N. De Jay, S. Papillon-Cavanagh, C. Olsen, N. El- Hachem, G. Bontempi, and B. Haibe-Kains, "mRMRe: an R package for parallelized mRMR ensemble feature selection". Bioinformatics, 29(18), 2365-2368, 2013.

[20] P. A. Mundra, and J. C. Rajapakse, "SVM-RFE with MRMR filter for gene selection", NanoBioscience, IEEE Transactions on, 9(1), 31-37, 2010.

[21] M. J. Garnett, E. J. Edelman, S. J. Heidorn, C. D. Greenman, A. Dastur, K. W. Lau, and Q. Liu, "Systematic identification of genomic markers of drug sensitivity in cancer cells", Nature, 483(7391), 570-575, 2012.

[22] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, and E. S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", science,286(5439), 531-537, 1999.

[23] M. Mandal, and A. Mukhopadhyay, "An improved minimum redundancy maximum relevance approach for feature selection in gene expression data", Procedia Technology, 10, 20-27, 2013.

[24] E. B. Huerta, B. Duval, and J. K. Hao, "A hybrid GA/SVM approach for gene selection and classification of microarray data", In Applications of Evolutionary Computing (pp. 34-44). Springer Berlin Heidelberg, 2006.

[25] O. Kurun, C. O. Akar, O. Favorov, N. Aydin, F. Urgen, "Using covariates for improving the minimum redundancy maximum relevance feature selection method", Turkish Journal of Electrical Engineering and Computer Sciences, 18(6):975–987, 2010.

[26] R. Li, X. Dong, C. Ma, and L. Liu, "Computational identification of surrogate genes for prostate cancer phases using machine learning and molecular network analysis", Theoretical Biology and Medical Modelling, 11(1), 37, 2014.

[27] X. Liu, A. Krishnan, and A. Mondry, "An entropy-based gene selection method for cancer classification using microarray data", BMC bioinformatics,6(1), 76, 2005.

[28] Y. Wang, F. S. Makedon, J. C. Ford, and J. Pearlman, "HykGene: a hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data", Bioinformatics, 21(8), 1530-1573, 2005.

[29] Soltani, M., Shammakhi, M. H., Khorram, S., & Sheikhzadeh, H. "Combined mRMR filter and sparse Bayesian classifier for analysis of gene expression data," In Signal Processing and Intelligent Systems (ICSPIS), International Conference. IEEE. pp. 1-5, Dec 2016.

[30] Chellamuthu, G., Kandasamy, P., & Kanagaraj, S., "Biomarker Selection from Gene Expression Data for Tumour Categorization Using Bat Algorithm. Methods," pp.402, June 2017.