Proceedings of the Third International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2019)
IEEE Xplore Part Number:CFP19OSV-ART; ISBN:978-1-7281-4365-1

Feature Selection in Cancer Genetics using Hybrid Soft Computing

S Thangavelu
Department of Computer Science and Engineering,
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
s_thangavel@cb.amrita.edu

Akshaya S
Department of Computer Science and Engineering,
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
aksh.sundar@gmail.com

K C Naetra
Department of Computer Science and Engineering,
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
kcnaetra@gmail.com

Krishna Sathya A C
Department of Computer Science and Engineering,
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
acssathya333@gmail.com

Vuyyuru Lasya
Department of Computer Science and Engineering,
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
vuyyurulasya1998@gmail.com

***Abstract*—Microarray databases are the most frequently used datasets for cancer analytics. Microarray databases are characterized by the presence of a very large number of genes, which exceeds the very little number of samples. So, the feature set accumulates the curse of dimensionality. Therefore, selecting a small subset of genes among thousands of genes in microarray data can potentially increase the accuracy for the classification of cancer. Many approaches, from the field of classical machine learning and soft computing, have been used to address the issue of feature selection and feature extraction for better classifications and clustering accuracy. The research outlined in this paper strives to look at a two-stage approach using minimum Redundancy Maximum Relevancy (mRMR), a feature ranking framework as the first stage followed by a hybrid genetic algorithm in the second stage that works on the features ranked by the mRMR. The proposed method is aimed to select the optimal feature subsets for better classification results in binary and multi class datasets to compensate for the curse of dimensionality in microarray datasets. The classifiers used to test the two-stage proposition are SVM, Naive-Bayes, Linear Discriminant Analysis, decision trees and random forest classifiers. The experimental results show that the gene subset selected by the mRMR-GA pipeline gives good results.**

Keywords—mRMR, microarray, Hybrid Genetic Algorithm, feature selection, two-stage, cancer classification, evolutionary computation, hybrid soft computing

I. INTRODUCTION

Input feature selection techniques are widely used to minimize irrelevant and redundant features from a large set of features. It is a necessary technique adopted when we have thousands of features because the presence of the irrelevant attributes could deviate the accuracy of a predictive model. Microarray databases represent the expression levels of thousands of genes in cancer tissues, resulting in the curse of dimensionality in it. Therefore, it calls for a feature selection approach to try and select the main genes responsible for a particular type of cancer. This paper explores a two-stage process to address the curse of dimensionality in microarray gene expression datasets. Given thousands of genes (gene markers) and limited number of samples, feature selection becomes a crucial step in accurate classification. Furthermore, the use of evolutionary algorithms like Genetic Algorithm would search for an optimised feature subset in the search space, taking feedback from the classification performance of the classifier being used. For performing the two-stage feature selection, mRMR (minimum Redundancy, Maximum Relevancy) is used as the first stage of feature selection. The top 'n' ranked features from the mRMR would then be passed to an evolutionary algorithm for the second stage of selection. The evolutionary algorithm would optimise the feature subset and select features accordingly using the mean squared error produced by the classifier as the fitness function.

II. RELATED WORK

Over the past few years, the development in the field of microarray technology has made it feasible to study the expression levels of tens of thousands of genes at the same time. Microarray datasets are characterized by many feature variables and relatively few observations. Microarray data analysis is aimed at hypothesis testing or hypothesis generation, which involves applications from a variety of computational disciplines like machine learning, deep learning, data mining, data science and soft computing.

Generally, microarray datasets are considered to be high-throughput biological data without loss of generality. In short, microarray technology is an experiment to provide large amounts of data that can dramatically enhance our understanding about biological activities [1]. DNA microarray technology measures and expresses the gene profile of thousands of genes. In biomolecular research, gene expression profiling is still a persisting issue. DNA microarrays measure a cell's transcript via the abundance of mRNA proteins [1]. There have been introductions of many hybrid approaches for gene expression experiments, which are believed to be a powerful tool for clinical diagnostics, as they help to uncover expression patterns that are characteristic of a particular disease. The fewer samples and explosive degree of feature attributes incur the curse of dimensionality. Therefore, feature selection and classification are two major tasks that accompany microarray gene expression data analysis along with other pre-processing strategies like missing values imputation and standardization [2].

Feature selection aims to find a subset of relevant features automatically or manually to reduce the dimensions of a structure and improve the prediction accuracy measured by using the selected features. Feature selection can also reduce the computation time and improve learning accuracy. There are three categories of feature selection methods namely the filter approach, the wrapper approach and the embedded method. In the filter approach, instead of removing the irrelevant and redundant features by using machine learning algorithms, the statistical techniques are used on the training data for ranking the significance of the gene subset or gene. In wrapper approach, the classifier is used to select those features that are best performing, by testing each of the selected features against a model. Besides wrappers and filters, there is an emerging technique called embedded methods which are a hybrid of the filter and wrapper methods. However, this embedded method is still in its infancy.

Stepwise Forward Selection (SFS) is one of the feature selection techniques, which follows the wrapper approach. According to this method, the predictive features are selected starting with single variable model followed by addition of each feature using a comparison criterion. The studies in [3] shows that GA performs better than SFS, irrespective of any classifier used. One drawback is that the GA based strategy works well by only taking the prediction accuracy into account and is not more elucidative from a biological perspective. Another study for selecting the predictive features in multiclass classification of tumor types was done by combining genetic algorithm (GA) and support vector machines (SVMs) followed by recursive feature elimination. Leave-one-out cross-validation (LOOCV) of the multiclass classification is given as the fitness test of GA [4]. Similarly, mRMR filter was used along with support vector machine - recursive feature elimination (SVM-RVE) for feature selection [5]. SVM-RVE is an iterative process used for ranking genes for cancer classification from gene expression datasets. The redundancy among the selected genes is reduced using this method.

In another paper [6], two-stage selection algorithm was used by combining mRMR and ReliefF. ReliefF, an attribute estimator is used to find the candidate gene set by providing dependencies between attributes. The second stage involves in selecting a compact subset of genes from the candidate set that has high relevance with the target class and also maximum dissimilarity within each other by applying mRMR. The application of ReliefF immediately followed by mRMR was able to overcome the inefficiency of ReliefF filter to isolate the irrelevant genes. Pipelining the selection process decreases the computational demand on the successive phases. A Dimension reduction of microarray was done using correlation-based filtering which filters the redundant genes on a particle swarm optimization search space [17]. This was followed by building a classification model using fuzzy rough quick reduct.

Initially hybrid soft computing techniques were aimed at selecting features and building the design of classification models concurrently [18]. One such work employs micro Genetic Algorithm (mGA) on features filtered by the Information Gain (IG) filter to achieve high classification accuracy with SVC [19]. Recent works deploy a three-phase scheme using mRMR-ABC and SVM classifier for selecting features in microarray datasets. The results suggested the promise of hybrid evolutionary approaches like mRMR-ABC in cancer gene selection and classification [20].

In this paper, a pipelining of the mRMR filter and Genetic Algorithm wrapper is accommodated, thereby presenting a two-stage selection algorithm. The mRMR method involves selection of genes that have the greatest bearing in context to the target label and are also highly unrelated to each other. Though GA is an effectual wrapper, it becomes computationally expensive when incorporated with a learning algorithm. The first phase of this algorithm involves the application of mRMR on the dataset to obtain a ranked potential candidate gene set. This results in filtering out many insignificant genes and reduction of the computational overload for the next phase. The second phase accommodates Genetic Algorithm that is directly exercised to pick out the optimal gene subset with the highest discrimination. The results convey that the optimal gene subset selected by the mRMR-GA pipeline improves the classification accuracy than when applied on the whole feature set.

III. PROPOSED WORK

This section explains the process flow and explores how the pre-processing, mRMR (minimum Redundancy Maximum Relevance) and the genetic algorithm pipeline works to identify important gene attributes.

A. System Architecture

The process flow is explained as represented in Fig. 1.

1. Three datasets are obtained from the database in the bioinformatics laboratories of Rutgers University [7]. The datasets are then subjected to pre-processing steps before feature selection.

2. The first pre-processing step is the missing values imputation. If the dataset has any missing values, then we go for local least squares imputation as demonstrated by [8].

3. Change of units in different microarray genetic tests makes it difficult to compare raw scores. So, Z-score transformation is performed on the values [9].

4. The pre-processed data is then given to mRMR to select arbitrary sets of features.

5. The selected features from the mRMR is given to the evolutionary algorithm that further forms a subset of features randomly. Different subsets lead to different candidates for the GA.

6. The classification accuracy, produced by a classifier, is used by the algorithm to measure and evaluate the fitness value for each candidate. Depending upon the classification accuracy for candidates, they are selected or discarded for the next generation.

7. Step 6 is followed for ‘n’ generations or until the MSE (Mean Squared Error) value reaches a threshold. Once

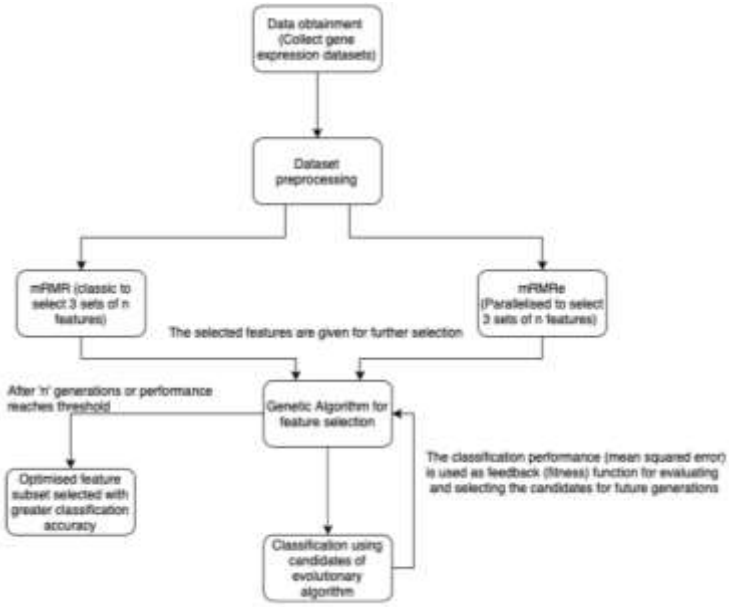

Fig. 1. System Architecture

either of the conditions is satisfied, the candidate with the highest fitness value would be considered as the optimized feature subset, selected using the two-stage implementation.

B. Data

The three datasets are collected from the database in the bioinformatics laboratories of Rutgers University [7]. Their specifications are as shown in table I and table II.

TABLE I.
OVERVIEW OF THE THREE DATASETS

Dataset Name	Number of Features (genes)	Number of Samples	Number of classes
Risinger Endometrial Cancer [14]	1771	42	4
Nutt-Brain Cancer [15]	1071	28	2
Pomeroy-Central Nervous System Embryonal Cancer [16]	88	34	2

TABLE II.
CLASSES AND SAMPLES

Risinger Endometrial Cancer [14]	Nutt Brain Cancer [15]	Pomeroy Central Nervous Embryonal Cancer [16]
Papillary serous Carcinoma (PS) (13)	Cancer Glioblastomas (CG)	Congenital Muscular Dystrophy (CMD) (25)
Clear cell carcinoma (CC) (3)	Non-cancer glioblastomas (NCG) (14)	Duchene Muscular dystrophy (DMD) (9)
Endometrial Carcinoma (E) (7)		
Normal Endometrium(N) (7)		

C. Pre-processing of datasets

Pre-processing is making the data ready for feature selection and classification. The first step is to check how many categorical and numeric variables are present in the dataset. The gene expression datasets are numeric datasets with the expression levels expressed in floating point numbers and only the target variable is categorical. Therefore, we can skip the dealing of categorical variables to convert them into numeric equivalents through Ordinal encoding or one hot encoding. The next step is to process the missing values, as they would have a major impact on the feature selection and classification. If the gene expression dataset has any missing values, then we would use the local least squares imputation, as this method has been proved to be competitive and appropriate for gene expression levels in microarray datasets [8].

Following the missing value implementation, standardization of the data is done by Z-score standardization. The Z-score for a given attribute is obtained by the linear transformation of the original value as shown in (1). The Z-scores for a particular attribute has a mean of zero and a standard deviation of 1 [10]. Moreover, Z-scores are famously used to compare raw scores that are taken from different tests [11]. Z-score complements really well with data that takes normal distribution. The motivation to prefer Z-score over other standardization techniques is supported by a competitive analysis of Z score transformation on microarray gene expression datasets conducted by Cheadle C et al. [22].

$$\bar{x} = \frac{x - \mu}{\sigma} \tag{1}$$

D. mRMR

The gene characteristics and its biological phenotypes are accurately identified and their relevance is narrowed down by using minimum redundancy feature selection algorithms. This is usually paired with relevant feature selection algorithm and is described as minimum Redundancy Maximum Relevancy (mRMR). The features that correlate strongest to the classification variable are selected using a scheme called as maximum-relevancy selection. On the contrary, there is another scheme termed as minimum-redundancy selection where the features can be selected in such a way that they are mutually far away from each other while still maintaining a high correlation to the classification variable. It has been found that mRMR selection turns out to be more powerful than the maximum relevance selection. As a special case, the statistical dependency between variables can replace the correlation. Mutual information can be used to quantify the dependency. In this case, the mRMR turns out to be an approximation to maximizing the dependency between the classification variable and the joint distribution of the features that are selected [12]. The mRMR starts functioning with an attribute that gives the maximal mutual information with the decision Y. Then, it greedily adds attribute X with a maximal value of the criterion measured by (2):

$$J(X) = I(X;Y) - \frac{1}{|S|}\sum_{W \in S} I(X;W) \tag{2}$$

Between numeric values, the Pearson coefficient is used to compute the mutual information between attributes. The reason for choosing the pearson coefficient over spearman coefficient because the data converged to the normal distribution after the application of Z-Score standardization. The mRMRe is a parallelized mRMR ensemble feature

selection. It is an extension of the classic mRMR technique by adding an ensemble approach to better explore the feature space.

The pre-processing and Z-score standardization is done using Python in Jupyter Notebooks. The mRMR libraries in R is used in the first stage of feature selection. Visualization of data was done using MATLAB.

The Risinger Endometrial cancer has no missing values. This is true for the other two datasets as well and hence the missing value imputation is skipped.

mRMR classic libraries is used in R to select the features. Since, the function requires the programmer to fix the number of features to select, the feature subsets have been selected robustly. The genetic algorithm would be run separately on each of the subsets and the results would be finally analyzed. The solutions method is used to select the fixed number of features (genes) using the feature count parameter, and store them in a 1D array. This array, containing the selected feature names, is written on to a separate CSV file. The selected features are given to the genetic algorithm for the second stage of selection.

E. Genetic Algorithm

The Genetic Algorithm (GA) follows the evolutionary algorithm approach. The special operators used in GA for the implementation are described in this section. The number of generations that the GA would go through is fixed robustly as 100.

1) Representation of Candidates

The candidates are represented as binary strings of length N, where N is the total size of features. If the feature is present in the candidate then the cell (gene) corresponding to that feature is 1, else if the feature is absent then that corresponding cell is 0.

2) Random initialisation

The feature subsets are randomly chosen from the feature set selected by the mRMR, at the beginning of the genetic algorithm.

3) Fitness

The fitness for each candidate is taken as the mean squared error (MSE), an important performance metric. MSE measures the deviation of the predicted values from the actual values. The classifier used to calculate this MSE is the Random Forest Classifier, which is an ensemble machine learning approach.

4) Crossover

One-point crossover is performed over the population. Two candidates are chosen based on tournament selection strategy. A random point for crossover is chosen between them and the crossover between the candidates is performed to generate two off - springs.

5) Mutation

Bit flipping is the most common mutation that is performed for candidates in GA. A random candidate is chosen for mutation. A random gene within the candidate is chosen and the bit is flipped. If 0, it is flipped to 1. If 1, it is flipped to 0. Mutation is done to achieve variability among the population.

6) Selection of parents for crossover

To select the fittest parents for crossover tournament selection strategy is used. There will be 9 tournaments of 5 randomly selected individuals. In each tournament, two individuals will be chosen (with likelihood of being chosen proportionate to their fitness) for crossover. The crossover will split them at a random spot and combine them.

7) Selection for next generation

The selection of candidates for the next generation is based on Elitism strategy. According to elitism, the top 20 fittest candidates will continue to the next generation.

IV. RESULT ANALYSIS

Different classical machine learning classifiers are used to compare the performance before and after the implementation of the proposed two stage feature selection. The results obtained after feature selection and before feature selection using different classifier are compared as shown in the Table 3, Table 4, and Table 5 and are illustrated in Fig. 2, 3 and 4.

As seen from the results classifier performance for the microarray datasets increases with the introduction of two stage feature selection. In medical predictive analytics, the classification accuracy is very important in deciding the fitness of a model. Along with hyper tuning of parameters, feature selection can also be done to increase the performance. One of the limitations is that the class balance and imbalance factor in the dataset affects the classifier performance. Naive Bayes and SVM perform well for Risinger- Endometrial and Nutt-Brain datasets with feature selection.

Another result worthy of discussion is the choice of 'N', where N is the final number of features that has been selected by the GA. There are many ways to determine the value of N. The most popular ways are Grid Search and iterative comparison [21-22]. In the proposed method, features had been ranked using mRMR filter (based on Pearson co-efficient) in the first stage. The top ranked features are given to the GA for feature selection. Initially, we had set each of the candidate's size to be 100. This means that each of the candidate is a binary string of length 100. The value 1 in the gene of a candidate refers to the feature being present and the value 0 refers to the feature being absent. Therefore, if for a given candidate 60 features are present then 60 cells correspond to 1 and remaining 40 cells correspond to 0. Each candidate is a subset of features. At the end of the 100th generation, the candidate with the highest fitness is selected as the final feature subset. In this implementation, after the 100th generation the candidate with the highest fitness has 50 cells corresponding to 1, this means that the winning candidate is a subset of 50 features. And hence, '50' selected features form the final feature subset, after the execution of the GA.

TABLE III.
PERFORMANCE COMPARISON FOR RISINGER ENDOMETRIAL DATASET

Classifier	Accuracy with 2-stage feature selection	Accuracy without 2-stage feature selection
Naive -Bayes - Gaussian	97.05%	96.6%
Random Forest Classifier	90%	76.6%
SVM	73.52%	73.3%

Fig. 2. Performance comparison for Risinger Endometrial Dataset

TABLE IV.
PERFORMANCE COMPARISON FOR POMEROY-CENTRAL NERVOUS SYSTEM EMBRYONAL CANCER

Classifier	Accuracy with 2-stage feature selection	Accuracy without 2-stage feature selection
SVM	91.6%	75%
Random Forest Classifier	87.5%	75%
LDA	97.9%	83.3%

Fig. 3. Performance comparison for Pomeroy-Central Nervous System Embroyonal Cancer Dataset

TABLE V.
PERFORMANCE COMPARISON FOR NUTT-BRAIN CANCER DATASET

Classifier	Accuracy with 2-stage feature selection	Accuracy without 2-stage feature selection
SVM	95%	85%
Random Forest Classifier	85%	70%
Naive Bayes	98.9%	94.6%

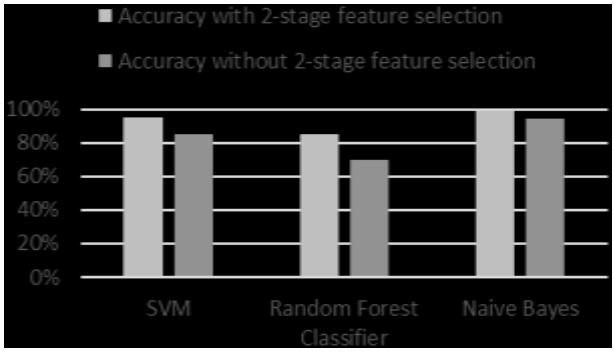

Fig. 4. Performance comparison for Nutt-Brain Cancer Dataset

V. CONCLUSION

The selection of classifiers for embedding in the GA has huge scope. Literature survey shows that classifiers ranging from Naive Bayes to Fuzzy Reduct systems have been used in accordance with GA for feature selection in multiple domains. A major challenge of our research work was the limited number of samples available in the standardized datasets chosen. This leads to the problem of over-fitting. Therefore, migration to bigger datasets with large number of samples would yield more well-grounded results. An improvement that could be done with huge microarray datasets is the incorporation of Neuro evolutionary strategies for feature selection. We had constrained ourselves to classical models because neural networks need huge samples of data for learning.

To enable this as a ground breaking research and realize market potential in the field of cancer genetics, we need to follow up with a genetics expert. Following up with a genetic expert or a biologist, can answer many questions like why the mRMR algorithm has ranked these genes as the top genes and analyse the correlation and mutual information matrix between the genes. A biological or genetic insight of such computational microarray data analysis was a major shortcoming in most of the literature works dealing with the same. If computational results could be coupled along with biological understanding then one could probably climb at least one step higher in understanding the genetics behind a disease and diagnosing it better.

REFERENCES

[1] Babu, M. Madan. "Introduction to microarray data analysis." *Computational genomics: Theory and application* 225 (2004): 249.

[2] Kumar, C. Arun, and S. Ramakrishnan. "Binary Classication of cancer microar- ray gene expression data using extreme learning machines." In 2014 *IEEE International Conference on Computational Intelligence and Computing Research*, pp. 1-4. IEEE, 2014.

[3] Luque-Baena, Rafael Marcos, Daniel Urda, Jose Luis Subirats, Leonardo Franco, and Jose M. Jerez. "Application of genetic algorithms and constructive neural networks for the analysis of microarray cancer data." *Theoretical Biology and Medical Modelling* 11, no. 1 (2014): S7.

[4] Peng, Sihua, Qianghua Xu, Xuefeng Bruce Ling, Xiaoning Peng, Wei Du, and Liangbiao Chen. "Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines." FEBS letters 555, no. 2 (2003): 358-362.

[5] Mundra, Piyushkumar A., and Jagath C. Rajapakse. "SVM-RFE with mRMR filter for gene selection."*IEEE transactions on nanobioscience* 9, no. 1 (2009):31-37.

[6] Zhang, Yi, Chris Ding, and Tao Li. "Gene selection algorithm by combining reliefF and mRMR." *BMC genomics* 9, no. 2 (2008): S27.

[7] Haznedar, Bulent. "Microarray Gene Expression Cancer Data." Mendeley Data. Mendeley Data, April 27, 2017. https://data.mendeley.com/datasets/ynp2tst2hh/4.

[8] Kim, Hyunsoo, Gene H. Golub, and Haesun Park. "Missing value estimation for DNA microarray gene expression data: local least squares imputation." *Bioin- formatics* 21, no. 2 (2004): 187-198.

[9] Cheadle, Chris, Marquis P. Vawter, William J. Freed, and Kevin G. Becker. "Analysis of microarray data using Z score transformation." *The Journal of molecular diagnostics* 5, no. 2 (2003): 73-81.

[10] R. G. van den Berg, "Z-Scores – What and Why?," SPSS tutorials. [Online]. Available: https://www.spss-tutorials.com/z-scores-what-and-why/. [Accessed: 27-Nov-2019].

[11] D. Statistics, "Z-score advantages and disadvantages- Akash Sarda," Z-score advantages and disadvantages- Akash Sarda, 01-Jan-1970. [Online]. Available: http://group1akash.blogspot.com/2013/10/z-score-advantages-and-disadvantages.html. [Accessed: 28-Nov-2019].

[12] "Minimum redundancy feature selection," Wikipedia, 27-Jul-2019. [Online]. Available: https://en.wikipedia.org/wiki/Minimum_redundancy_feature_selection. [Accessed: 29-Nov-2019].

[13] "Which correlation coefficient is better to use: Spearman ..." [Online]. Available: https://www.researchgate.net/post/Which_correlation_coefficient_is_better_to_use_Spearman_or_Pearson. [Accessed: 29-Nov-2019].

[14] "Risinger_Endometrial Cancer," Mendeley Data - Viewer - Risinger_Endometrial Cancer.xlsx. [Online]. Available: https://data.mendeley.com/datasets/ynp2tst2hh/2/files/404aec5a-07e2-4e8e-b384-3e9f6b039be6. [Accessed: 29-Nov-2019].

[15] "Nutt-2003-v2_BrainCancer," Mendeley Data - Viewer - Nutt-2003-v2_BrainCancer.xlsx. [Online]. Available: https://data.mendeley.com/datasets/ynp2tst2hh/4/files/8f5ed0a3-f4cd-47a8-88dc-287df5c8ab9b. [Accessed: 29-Nov-2019].

[16] "Pomeroy-2002-v1_CentralNervousSystemEmbryonalCancer," Mendeley Data - Viewer - Pomeroy-2002-v1_CentralNervousSystemEmbryonalCancer.xlsx. [Online]. Available:https://data.mendeley.com/datasets/ynp2tst2hh/4/files/29d88b74-e4ed-4092-9cff-617d139fe924. [Accessed: 29-Nov-2019].

[17] Arunkumar, C., and S. Ramakrishnan. "A hybrid approach to feature selection using correlation coefficient and fuzzy rough quick reduct algorithm applied to cancer microarray data." In 2016 *10th International Conference on Intelligent Systems and Control* (ISCO), pp. 1-6. IEEE, 2016.

[18] Saad, Ashraf. "An overview of hybrid soft computing techniques for classifier design and feature selection." In 2008 *Eighth International Conference on Hybrid Intelligent Systems*, pp. 579-583. IEEE, 2008.

[19] Pragadeesh, C., Rohana Jeyaraj, K. Siranjeevi, R. Abishek, and G. Jeyakumar. "Hybrid feature selection using micro genetic algorithm on microarray gene expression data." *Journal of Intelligent & Fuzzy Systems* 36, no. 3 (2019): 2241-2246.

[20] Alshamlan, Hala, Ghada Badr, and Yousef Alohali. "mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling." *Biomed research international* 2015 (2015).

[21] Ferroudji, Karim, Bahia Yahya-Zoubir, Maouia Bentlemsan, and ET-Tahir Zemouri. "Features Selection Using Differential Evolution in Motor-Imagery Based Brain Machine Interface." In *Proceedings of the International Conference on Intelligent Information Processing, Security and Advanced Communication*, p. 84. ACM, 2015.

[22] Cheadle, Chris, Marquis P. Vawter, William J. Freed, and Kevin G. Becker. "Analysis of microarray data using Z score transformation." *The Journal of molecular diagnostics* 5, no. 2 (2003): 73-81.