DOI: https://doi.org/10.52756/ijerr.2024.v39spl.006 *Int. J. Exp. Res. Rev.*, Vol. 39: 82-91 (2024)

Original Article | Peer Reviewed | Open Access

International Journal of Experimental Research and Review (IJERR)
© Copyright by International Academic Publishing House (IAPH)
ISSN: 2455-4855 (Online)
www.iaph.in

Feature Selection in Microarray using Proposed Hybrid Minimum Redundancy-Maximum Relevance (MRMR) and Modified Genetic Algorithm (MGA)

P. Nancy Vincentina Mary[1,2]* and R. Nagarajan[2]

[1]Department of MCA, Fatima College, Madurai, Tamil Nadu, India; [2]Department of Computer and Information Science, Faculty of Science, Annamalai University, Chidambaram, Tamil Nadu, India

E-mail/Orcid Id:

PN, nancy.vincentina.mary@gmail.com, https://orcid.org/0000-0003-0677-5483;
RN, rathinanagarajan@gmail.com, https://orcid.org/0000-0001-7733-5085

Article History:

Received: 02nd Jan., 2024
Accepted: 17th May, 2024
Published: 30th May, 2024

Keywords:
Classification, Feature Selection, Gene Selection, Gene Expression Classification, Microarray Classification, Modified Genetic Algorithm

How to cite this Article:
P. Nancy Vincentina Mary and R. Nagarajan (2024). Feature Selection in Microarray using Proposed Hybrid Minimum Redundancy-Maximum Relevance (MRMR) and Modified Genetic Algorithm (MGA). *International Journal of Experimental Research and Review*, *39*(spl.) 82-91.
DOI:
https://doi.org/10.52756/ijerr.2024.v39spl.006

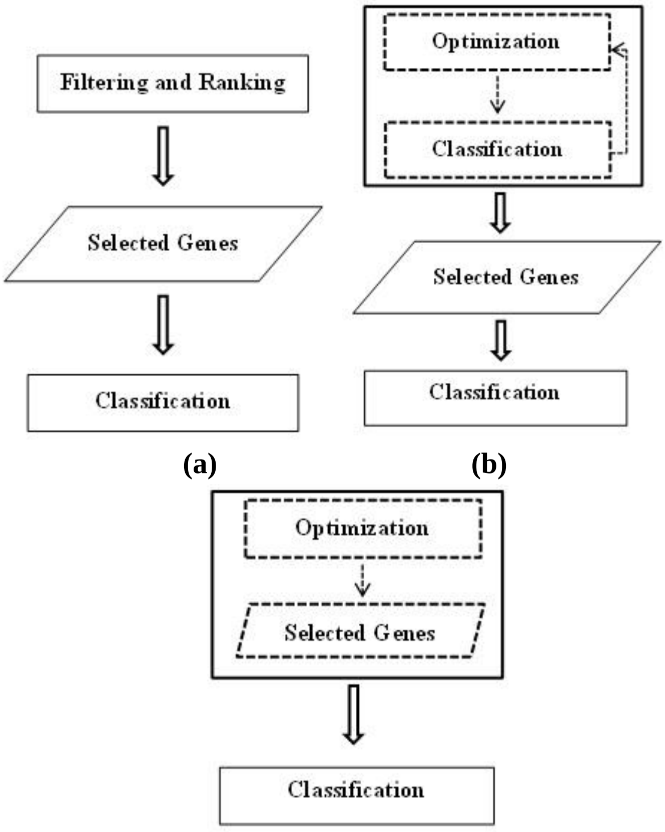
Abstract: Gene expression microarray data commonly have an enormous number of genes with a smaller number of samples. In these genes, many are irrelevant, insignificant or redundant for the classification analysis. Therefore, the identification of informative genes, which have the greatest role in classification and diagnosis, is of essential and practical importance to the classification problems, such as cancer versus non-cancer classification and classification of different tumor types. This paper aims to present a novel idea for implementing MRMR, the hybrid Minimum Redundancy-Maximum Relevance method combined with a Modified Genetic Algorithm, to minimize the selection of microarray data feature sets. This paper proposes a two-step feature selection algorithm by integrating Minimum Redundancy Maximum Relevance (MRMR) and Modified Genetic Algorithm (MGA). In the first step, MRMR is used to filter redundant genes in high-dimensional microarray data. The second step is used to eliminate irrelevant genes. The proposed MRMR-MGA algorithm is compared with traditional MRMR with the GA algorithm. The implementation results show that the proposed method has good selection and classification performances.

Introduction

Recent developments in microarray technology have made it possible to analyze thousands or more than thousands of genes simultaneously. However, the major problem in this analysis is the huge amount of genes compared to the limited amount of samples (Hajieskandar et al., 2023). Most classification algorithms suffer from such a high-dimensional input space. Furthermore, many of the genes in arrays are irrelevant or redundant to some specified diseases. Thus, selecting highly discriminating genes is critical to improving the accuracy of disease classification and prediction (Shukla et al., 2020; Mishra et al., 2023). Identification of a set of genes that suitably differentiate biological samples of various types is a feature selection problem. Feature selection includes finding a subset of features to improve prediction and classification accuracy or decrease the size of the data with only the selected features without decreasing the prediction and classification of the classifier (Zare et al., 2023).

Methods for feature selection are generally divided into three categories: the filter approach, the wrapper approach, and the embedded method. In the first category, a filtering approach is first used to choose a subset of features before applying the actual model learning algorithm. Features were selected based on their scores on various statistical tests of correlation with the outcome variable. Filter Methods (Almugren et al., 2019) mainly act as rankers, ordering the features from best to worst. The ranking of features depends on the essential

*Corresponding Author: nancy.vincentina.mary@gmail.com properties of the data, such as variance, consistency, distance, information, correlation, etc. On the other hand, the wrapper approach (Osama et al., 2023) utilizes the learning machine as a fitness function and searches for the best feature subset in the space of all feature subsets. In wrapper methods, feature subsets are used to train a model. Based on the inferences from the trained model, features are added or eliminated from the subset. In other words, Wrapping methods compute a model with a specific feature subset and estimate the importance of each feature. It then iterates and tries different subsets of the feature until it reaches the optimal subset. Besides wrappers and filters, the embedded methods (Liu et al., 2018) are another category of feature selection algorithms, which perform feature selection as the process of training and are usually specific to given learning machines. Embedded methods bridge the gap between filters and wrappers. To begin with, this method fuses measurable and statistical criteria like a filter to choose some features, and then using a machine learning algorithm, this method picks the subset with the best classification performance. Figure 1, presents a logical diagram to show the relationship between filter, embedded, and wrapper approaches (Syahidin et al., 2023).



(a) (b)

(c)

Figure 1. Different feature selection techniques (a) filter (b) wrapper (c) embedded.

All feature selection techniques need an evaluation function and a search strategy to obtain the optimal feature set. The evaluation function tries to measure the discriminating ability of a feature or of a subset to distinguish the different class labels and can be grouped into five categories (Bhartiya and Prajapati, 2023) they are distance, uncertainty, dependence, consistency and classifier error. Searching for the optimal subset can be achieved by examining all possible subsets, which is usually unfeasible in practice due to the large amount of computational effort required. A wide range of heuristic search strategies are used, including forward selection, backward elimination, hill-climbing, branch and bound algorithms, and stochastic algorithms like simulated annealing and GA.

Genetic algorithms appear to show a lot of promise when they are used in the medical diagnosis of various diseases in terms of promoting more exact predictions as well as better-tailored treatment regimens (Tyagi et al., 2024; Ghaheri et al., 2015; Balcha and Woldie, 2023). Various feature selection methods have been used in gene selection for cancer classification. Minimum Redundancy Maximum Relevance (MRMR) is used to find the optimal subset of multiple genes (Alromema et al., 2023). The features of the MRMR method are extremely different from each other. The mutual Euclidean distance is maximized or the pairwise correlation is minimized. The usual maximum association criteria, such as maximum mutual information with the target phenotype, supplement these minimum redundancy criteria. This filtering technique has been proven to achieve high accuracy by eliminating irrelevant and redundant features.

Unlike filters, wrappers use the estimated accuracy of a specific classifier to evaluate candidate subsets. Genetic algorithms have been utilized to realize dimensionality reduction. In a GA-based wrapper approach (El Akadi et al., 2011), each feature is represented as a gene and a feature subset as a chromosome. The occurrence or non-occurrence of a feature corresponds to a chromosome value of 1 or 0. First, a population of chromosomes is randomly created. All chromosomes are evaluated using a fitness function to determine their fitness value. Chromosomes with higher fitness are retained, chromosomes with lower fitness are discarded, and new populations are created through crossover and mutation. These operations are designed to make the next generation healthier. As this process repeats, the chromosome population evolves, expecting stronger chromosomes to emerge and survive. In most real-world situations, the chromosomes will improve and become

DOI: https://doi.org/10.52756/ijerr.2024.v39spl.006

more appropriate. At the end of the GA, the archived gene subset is examined. The gene subsets are compared, and the highest fitness value corresponding to the gene subset is selected. This process usually leads to further reduction of the gene set.

Materials and Methods

In microarray data, selecting a few important genes allows efficient data analysis and supports its biological interpretation. This section describes the MRMR filtering method and a Modified GA-based wrapper method for gene selection or feature selection. These algorithms are developed to obtain gene subsets as a solution to decrease the large number of genes that need to be classified later. The MRMR method is utilized to select genes that are most related to the target class and extremely different from each other. Modified GA is a general and efficient wrapper. Combining MRMR and MGA can give an effective gene selection method. In this proposed method, gene selection is performed in two stages. In the first stage, MRMR is used to find a candidate gene set. MRMR removes unimportant genes. In the second stage, the MGA method is applied to select the highest discriminative gene subset from the candidate set which is obtained using MRMR. The microarray dataset of ovarian cancer in CSV format (Hameed et al., 2021) is used. After the feature selection method, a Random forest classifier is used to classify the dataset. This paper uses a Random Forest classifier to compare the MRMR-MGA feature selection algorithm with the MRMR-GA. The implementation results show that MRMR-MGA gene selection is effective compared to the MRMR-GA algorithm.

MRMR Filter Method

The mutual information between two variables X and Y, denoted I (X; Y), is a quantity that measures the mutual dependence of these two variables (El Akadi et al., 2011). The mutualinformation between two discrete variables X and Y can be formulated as follows.

$$I\ (X;\ Y\ ) = \sum_{y\in Y} \cdot \sum_{x\in X} \cdot p(x;\ y) \log \frac{p(x;\ y)}{p(x)p(y)} \quad \text{-------} \qquad (1)$$

Here the joint probability distribution function of *X* and *Y*is is mentioned as *p(x, y)*, and the marginal probability distribution functions of *X* and *Y* are mentioned as *p(x)* and *p(y)*, respectively. For continuous value, the summation is replaced by a double integral. Naturally, mutual information measures the information that *X* and *Y* share. It measures the extent to which knowledge about one of these variables reduces uncertainty about the other variable. Using the concept of mutual information, the MRMR method (Mandal et al., 2013) selects genes that have the maximum relevance and minimum redundancy with the target class. In other words, the selected genes are extremely dissimilar to each other. Given g_i, which represents the gene *i*, g_j, which denotes the gene *j*, and the class label *c*, the top *m* genes are carefully chosen using the Maximum-Relevance method in the descending order of $I(g_i;\ c)$, i.e. c class labels related to the finest *m* individual features.

$$\max S \frac{1}{|S|} \sum_{g_i \in S} I(g_i;\ c) \quad \text{-------} \qquad (2)$$

Even though the topmost individual genes are selected using the Maximum-Relevance algorithm, it has been accepted that the *m* finest features are not the best *m* features, since the associations among those topmost features may also be high. A minimum redundancy method is used to remove the redundant features.

$$\min S \frac{1}{|S|^2} \sum_{g_i g_{j,} \in S} I\ (g_i;\ g_j) \quad \text{-------} \qquad (3)$$

The minimum redundancy–maximum relevance (MRMR) feature selection method combines both optimization criteria of equations 2 and 3. A sequential incremental algorithm to solve the simultaneous optimizations of optimization criteria of equations 2 and 3 is given as follows. Suppose a Set G represents the set of genes; for a given Sm−1, the feature is set with m−1 genes. Then, the aim is to select the feature from the Set G – S_{m-1}. This attribute is selected by maximizing the single-variable relevance and subtracting a redundant function.

$$\max g_i \in \mathrm{G} - \mathrm{S}_{m-1}\Big(I(g_i;\ c) - \frac{1}{m-1}\Big(\sum_{g_{j,} \in S_{m-1}} I\ (g_i;\ g_j)\Big)\Big) \text{-------} \qquad (4)$$

Steps involved in MRMR:

1. **Step 1:** (Relevance Calculation) For each feature, calculate its relevance to the target variable. This is calculated using Mutual Information.
2. **Step 2:** (Redundancy Calculation) Their redundancy is calculated for each pair of features.
3. **Step 3:** (Initialize Lists): Initialize two lists: one for selected features and another for remaining candidate features
4. **Step 4:** (Select Initial Feature): Choose the feature with the highest relevance as the first selected feature.
5. **Step 5:** (Iterative Feature Selection): Repeat the following steps until the desired number of features is selected. Select the feature with the maximum MRMR criterion and add it to the list of selected features. Remove the selected feature from the list of candidate features. The final selected features are stored in a list

DOI: https://doi.org/10.52756/ijerr.2024.v39spl.006

GA Wrapper Method

Gene expression data often contains redundant, irrelevant, and noisy genes. The presence of such genes during the learning process of machine learning algorithms impacts the performance. The performance will affect the computational cost and prediction accuracy. Gene selection is the process of reducing the dimensionality of a dataset and identifying a small set of biologically relevant genes to achieve classification results comparable to or better than using all genes (Li et al., 2013). Wrapper methods are frequently used in feature selection (gene selection) for gene expression analysis in cancer prediction or cancer classification to distinguish between tumour types, decrease the number of genes tested in new patients, and also aid in drug discovery and early diagnosis. Genetic algorithms (Albadr et al., 2020) are meta-heuristic algorithms based on the process of natural selection to obtain optimal and high-quality solutions for producing offspring through the application of genetic operators, namely selection, crossover, and mutation. The set of possible solutions to a given problem is called the initial population, and each member of the population is called a chromosome.

Chromosomes are made up of collections of genes, and all chromosomes contain the same number and type of genes. As part of the optimization process, several genes are selected from the population and genetically bred through crossover and mutation to obtain the next generation of offspring with better fitness. Crossover and mutation are two operators used to generate new populations. In the crossover process, two chromosomes (parents) are joined to form a new chromosome called the offspring. Crossover operators are applied repeatedly. Genes with good chromosomes appear in the population. Crossovers are frequently applied in GA. Mutations play an important role in GA. Applying mutation operators causes random changes in the properties of chromosomes. Many generations of GA runs are required until the solution with maximum fitness is obtained and returned as the best solution to the problem.

Steps involved in GA:

1. **Individual encoding:** Each individual is encoded as a binary vector of size, where the entry $b_i = 1$ represents the predictor $p_i$that is defined for that individual, $b_i = 0$ if the predictor p_i is not included in that particular individual ($i = 1, \ldots, P$) (Cerradaet al., 2016).
2. **Initial population:** Given a binary representation of individuals, the population is a binary matrix where the rows are randomly selected individuals and the columns are the available predictors. An initial population containing a predefined number of individuals is generated using a random selection of 0 and 1 for each entry (Cerrada et al., 2016).
3. **Fitness function:** The fitness value of each individual in the population is calculated using a predefined fitness function (Welikala et al., 2015). This fitness function selects individuals for the next generation, which has the lowest prediction error and fewer predictors.
4. Create the next generation by applying genetic operators.
   - **Selection:** Elite individuals are selected based on their fitness values. They are selected as parents to produce children through a process of crossover and mutation. Instead of selecting all parents from the most eligible individuals, this study adds random individuals to the parent pool to maintain generational diversity. Each pair of parents produces some children to form the next generation of the same size as the original population. To stabilize the size of each generation, equation (5) must be satisfied.

$$\frac{of\ BS + \#of\ RS}{2}\ X\quad of\ children = initial\ population\ size\ \text{-------} \tag{5}$$

   Here Best Selected Individuals are represented as BS and Randomly Selected Individuals are represented as RS.

   - **Crossover:** A mechanism by which a new generation is created by exchanging entries between the two parents selected in the previous step. This study uses a single-point crossover technique (Liu et al., 2013; Welikala et al., 2015).
   - **Mutation:** This operation is applied after crossover to determine whether an individual should be mutated in the next generation, ensuring that the predictor has not been permanently removed from the GA population (Brown and Sumchrast, 2005).
5. **Stopping Criteria:** Two stopping criteria are commonly used in GA. The first method is used to reach the maximum number of generations. Another one is that the fitness function does not improve in two consecutive generations (Cheng et al., 2016). Steps 2 and 3 are executed repeatedly until the stopping criteria are met.

Modified Genetic Algorithm (MGA) Wrapper Method

The modified GA aims to direct the stochastic selection aspect towards a fine subset of features. This modified genetic algorithm is used to find the best combination of features. Using a genetic algorithm for gene expression microarray feature selection involves applying principles inspired by natural selection and

DOI: https://doi.org/10.52756/ijerr.2024.v39spl.006

evolution to iteratively search for an optimal subset of genes relevant to a specific task.

In the Modified Genetic Algorithm, modification is done in two phases. One is in the selection process, instead of selecting random features, features are ranked and the top features are selected. The second modification is in the crossover phase. In the evolutionary genetic algorithm AND/ OR operators are used for crossover. Information Gain-based features are combined in the crossover phase.

Proposed hybrid feature selection in Microarray using Minimum Redundancy-Maximum Relevance (MRMR) and Modified Genetic Algorithm (MGA)

The work reported in this paper is based on a hybrid approach combining MRMR and MGA. The proposed method is characterized by two stages: In the first stage, MRMR is used to filter noisy and redundant genes in high-dimensional microarray data. In the second stage, MGA for selecting the highly discriminating genes. The scheme for the proposed model is shown in Figure 2. The detailed step is described as follows. In the first stage, the original data are preprocessed by the MRMR filter. Each gene is evaluated and sorted according to the MRMR criterion, and the first P genes are selected to form a new subset. MRMR is applied to filter out many important genes and reduce the computational load. The second stage uses a wrapper approach MGA to accomplish the feature(gene) subset selection from the reduced set of genes obtained in the previous pre-processing stage. The basic idea here consists of using a GA to discover good subsets of genes. Modification is done in the selection part and crossover part. In this modified method number of iterations is increased in the selection base and features are selected using rank. In the crossover phase, the evolutionary algorithm used the AND/OR function but this proposed method uses Information gain, the goodness of a subset being evaluated by a Random Forest classifier.

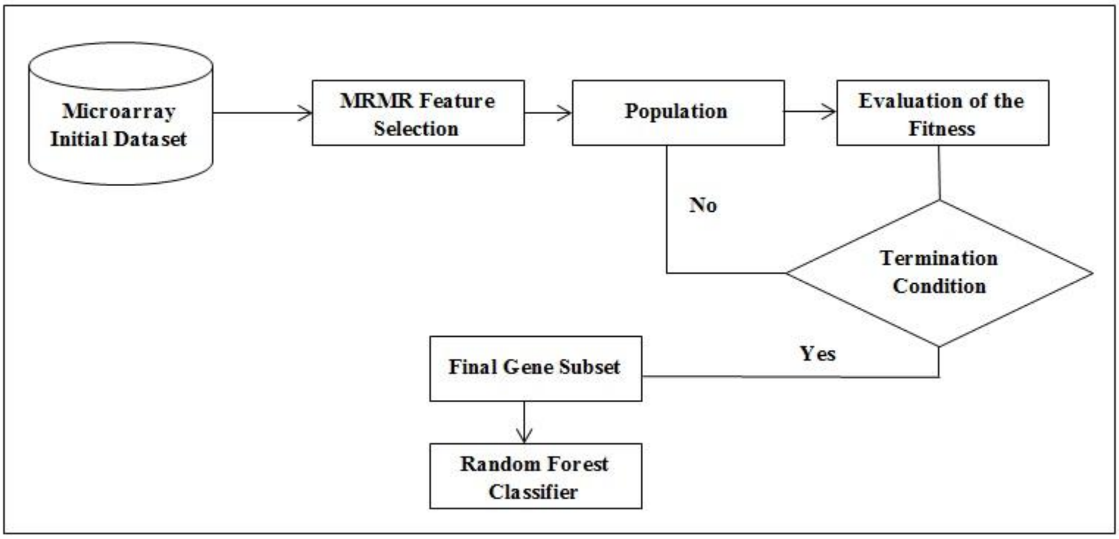


Figure 2. The general process for gene selection and classification using the MRMR-MGA algorithm.

Stages involved in MGA:

1. **Initialization:** Create an initial population of feature subsets.
2. **Fitness Function:** Define a fitness function to evaluate the performance of each feature subset.
3. **Selection:** Select individuals (feature subsets) from the current population based on their fitness. Rank-based features are selected in this phase.
4. **Crossover (Recombination):** Apply crossover (recombination) to pairs of selected individuals to create new feature subsets. In this phase, information gain is used to cross over.
5. **Mutation:** Introduce random changes to some individuals in the population.
6. **Evaluate Fitness of Offspring:** Evaluate the fitness of the newly created individuals (offspring) using the fitness function.
7. **Replacement:** Replace some individuals in the current population with newly created individuals.

Repeat the steps 3 – 7 until the desired features are selected. The algorithm returns the feature subset with the highest fitness in the final population as the selected set of genes for the gene expression microarray data.

Random Forest Classifier for Classification

Random Forests (RF) is a classification algorithm that uses an ensemble of unpruned decision trees based on bootstrap samples of training data with a randomly selected subset of variables. This algorithm has many

DOI: https://doi.org/10.52756/ijerr.2024.v39spl.006

characteristics that make it an attractive technique in the classification of microarray gene expression data. Random Forest (RF) is a tree-based method that proposes a modified approach to bagging techniques to construct non-correlated tree collections. T_b, b=1,…, B, with low bias (low error on the training data) and low variance (low error on the test data) by averaging their predictions (Breiman, 2001). Reducing variance by reducing the correlation between trees is achieved by random selection of input variables and by replacing samples from a dataset of size m. The selected variables and samples are used to grow each tree in the forest (bootstrap sampling). This random selection has shown that around 2/3 of the data are chosen, then, the training set m_b for each classifier is, in general, $m_b \subset m$ (Ziegler et al., 2014).

Steps involved in Random Forest Classification:

1. Uses the test features and uses the rules of every randomly created decision tree to predict the outcome. Then it stores the predicted outcome(target)
2. The votes for each predicted target are calculated.
3. Take the high-voted predicted target as the final prediction from the random forest algorithm

Results and Discussion

This section performs comprehensive implementation to compare the MRMR-MGA feature selection algorithm with MRMR-GA using the Random Forest Classifier on the ovarian cancer dataset. This paper used an Ovarian Cancer Data set with 253 records and 15154 Features. MRMR technique is applied to the whole Dataset, and 200 Features are selected from this method. A dataset with 200 features is applied to GA and 45 features are selected. A dataset with 45 features is classified with a Random Forest classifier. MRMR and GA methods are implemented using Python, and classification is performed using the WEKA tool. When using the proposed method, MRMR technique is applied to Full Dataset and 200 Features are selected from this method. A dataset with 200 features is applied to Modified GA and 21 features are selected. Datasets with 21 features are classified with a Random Forest classifier. MRMR and GA methods are implemented using Python, and classification is performed using the WEKA tool.

The classification accuracy of three different datasets is tested using a Random Forest classifier.

1) Original Data set with 15154 features are directly applied to the Random Forest Classifier
2) The MRMR technique is applied to the original dataset in the second method. This filter method removes redundant features and gives 200 features. Then GA method is applied to this reduced dataset. This wrapper method removes irrelevant features and gives 45 features. Datasets with 45 features are applied to the Random Forest Classifier.
3) The MRMR technique is applied to the original dataset in the third method. This filter method removes redundant features and gives 200 features. Then, a Modified Genetic Algorithm is applied to this reduced dataset. This wrapper method removes irrelevant features and gives 21 features. Datasets with 21 features are applied to the Random Forest Classifier.

Table 1 describes the classification accuracy of three methods. One is the classification result without feature selection, the second one is the classification result with MRMR with GA-based feature selection and the third one is the MRMR with modified GA-based feature selection.

The classification results of the Original Data set with 15154 features using the Random Forest classifier are shown in Figure 3.

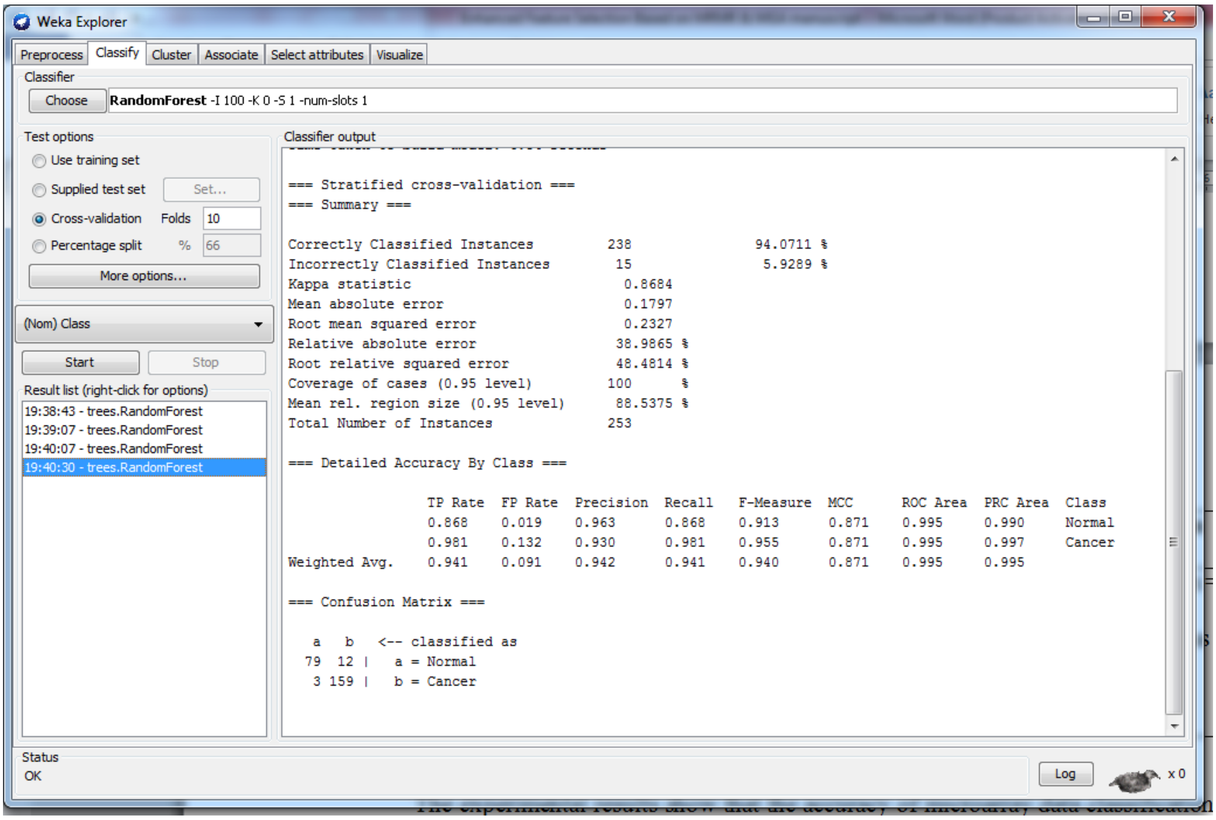
The classification results of the Data set with MRMR-GA-based feature selection and with 45 features using the Random Forest classifier are shown in Figure 4.

The classification results of the Data set with MRMR-Modified GA-based feature selection and with 21 features using Random Forest classifier are shown in Figure 5.

The Mean Absolute Error of the classifier when using these three datasets are shown in Figure 6. The mean absolute Error of the classifier when using the original dataset is 1.1797. MRMR- GA-based feature selection is applied to the original dataset, the classifier error rate is 0.0462. The error rate is 0.0183 when MRMR-MGA is applied to the original dataset.

Table 1. Accuracy of Classifiers.

Classification Result (Without Feature Selection)	Classification Result (With MRMR + GA Feature Selection)	Classification Result (With MRMR + Modified GA Feature Selection)
=== Confusion Matrix === a b <-- classified as 79 12 | a = Normal 3 159 | b = Cancer	=== Confusion Matrix === a b <-- classified as 87 4 | a = Normal 2160 | b = Cancer	=== Confusion Matrix === a b <-- classified as 91 0 | a = Normal 0 162 | b = Cancer



Classifier: RandomForest -I 100 -K 0 -S 1 -num-slots 1

● Cross-validation Folds 10

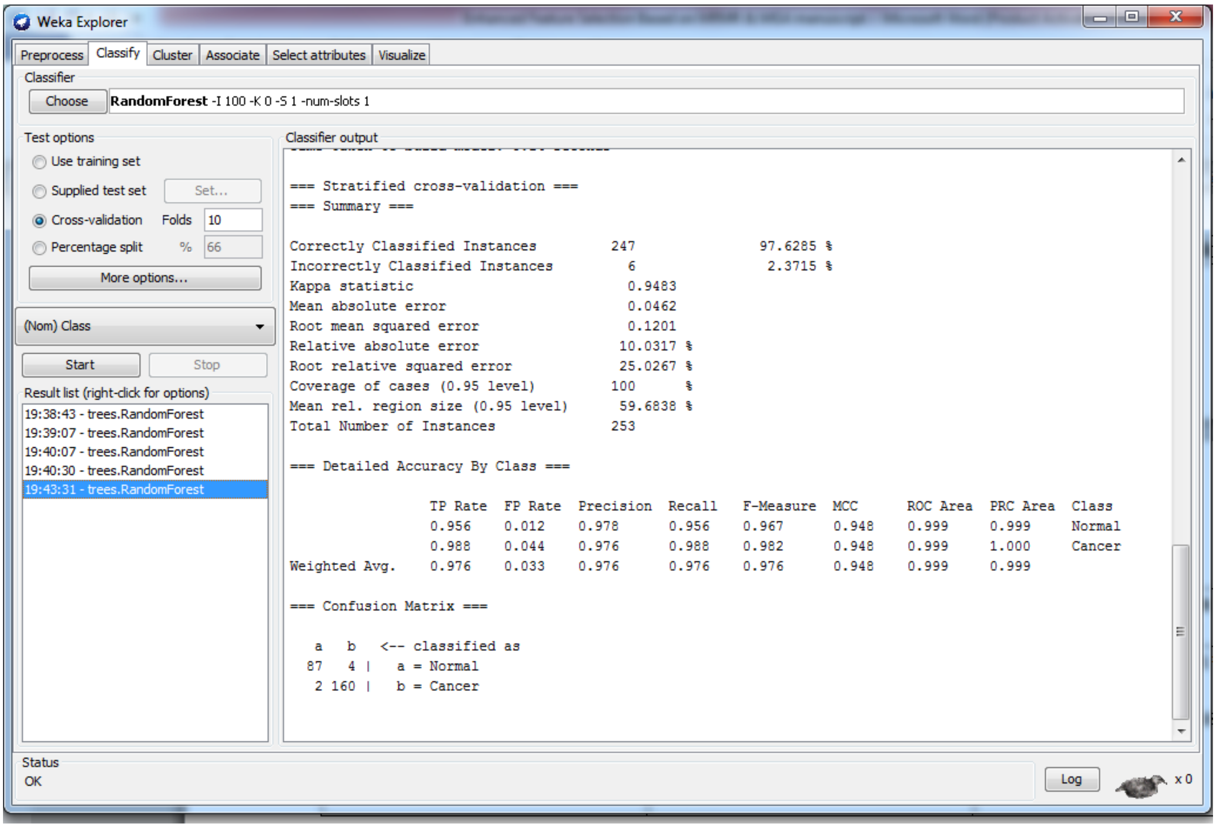
(Nom) Class

19:38:43 - trees.RandomForest
19:39:07 - trees.RandomForest
19:40:07 - trees.RandomForest
19:40:30 - trees.RandomForest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         238               94.0711 %
Incorrectly Classified Instances        15                5.9289 %
Kappa statistic                          0.8684
Mean absolute error                      0.1797
Root mean squared error                  0.2327
Relative absolute error                 38.9865 %
Root relative squared error             48.4814 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      88.5375 %
Total Number of Instances              253

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.868    0.019    0.963      0.868    0.913      0.871    0.995     0.990     Normal
                 0.981    0.132    0.930      0.981    0.955      0.871    0.995     0.997     Cancer
Weighted Avg.    0.941    0.091    0.942      0.941    0.940      0.871    0.995     0.995

=== Confusion Matrix ===

   a   b   <-- classified as
  79  12 |   a = Normal
   3 159 |   b = Cancer
```

Status: OK

Figure 3. Classification results of the Original Data set.



Classifier: RandomForest -I 100 -K 0 -S 1 -num-slots 1

● Cross-validation Folds 10

(Nom) Class

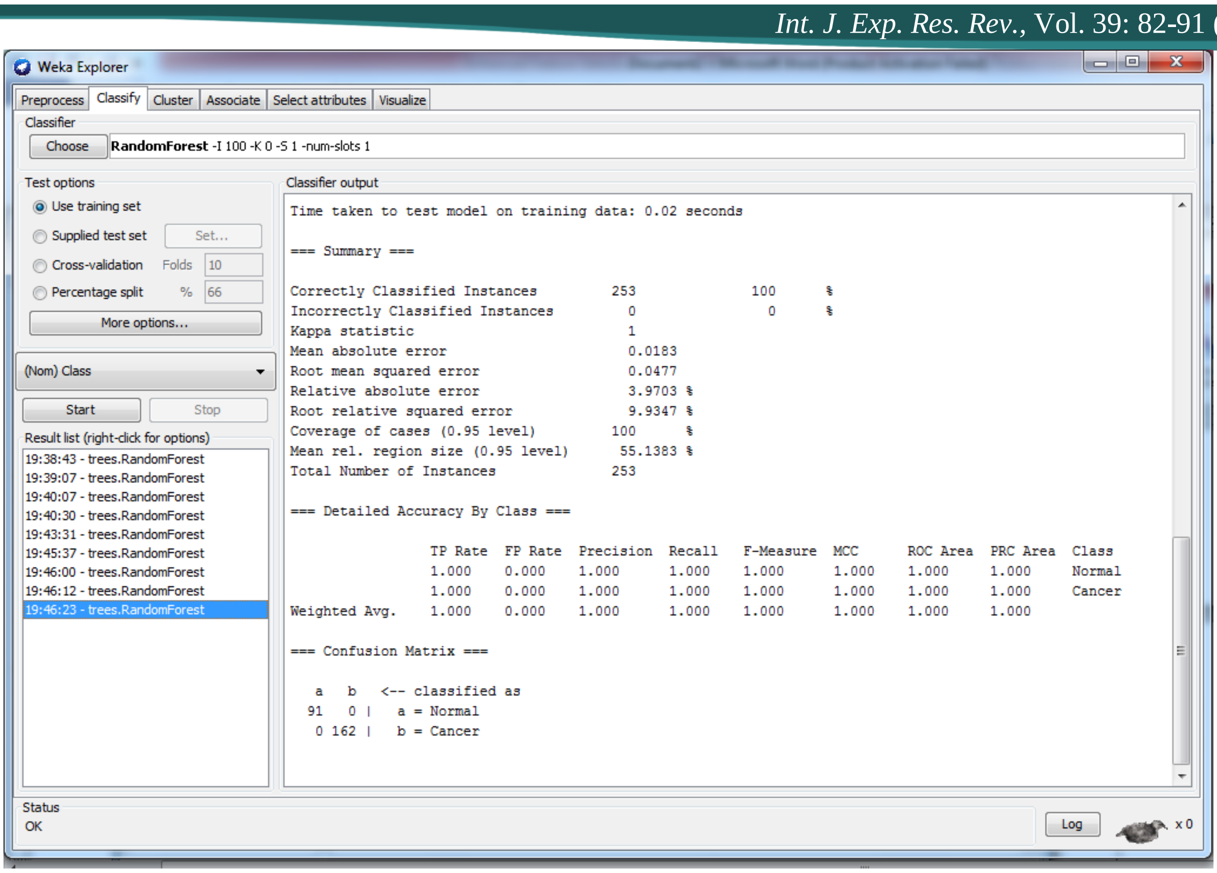
19:38:43 - trees.RandomForest
19:39:07 - trees.RandomForest
19:40:07 - trees.RandomForest
19:40:30 - trees.RandomForest
19:43:31 - trees.RandomForest

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         247               97.6285 %
Incorrectly Classified Instances         6                2.3715 %
Kappa statistic                          0.9483
Mean absolute error                      0.0462
Root mean squared error                  0.1201
Relative absolute error                 10.0317 %
Root relative squared error             25.0267 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      59.6838 %
Total Number of Instances              253

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.956    0.012    0.978      0.956    0.967      0.948    0.999     0.999     Normal
                 0.988    0.044    0.976      0.988    0.982      0.948    0.999     1.000     Cancer
Weighted Avg.    0.976    0.033    0.976      0.976    0.976      0.948    0.999     0.999

=== Confusion Matrix ===

   a   b   <-- classified as
  87   4 |   a = Normal
   2 160 |   b = Cancer
```

Status: OK

Figure 4. Classification results of MRMR – GA applied Data set.

DOI: https://doi.org/10.52756/ijerr.2024.v39spl.006



Classifier: RandomForest -I 100 -K 0 -S 1 -num-slots 1

Test options: ● Use training set ○ Supplied test set ○ Cross-validation Folds 10 ○ Percentage split % 66

(Nom) Class

```
19:38:43 - trees.RandomForest
19:39:07 - trees.RandomForest
19:40:07 - trees.RandomForest
19:40:30 - trees.RandomForest
19:43:31 - trees.RandomForest
19:45:37 - trees.RandomForest
19:46:00 - trees.RandomForest
19:46:12 - trees.RandomForest
19:46:23 - trees.RandomForest
```

```
Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correctly Classified Instances         253               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0.0183
Root mean squared error                  0.0477
Relative absolute error                  3.9703 %
Root relative squared error              9.9347 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      55.1383 %
Total Number of Instances              253

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Normal
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Cancer
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
  91   0 |   a = Normal
   0 162 |   b = Cancer
```

Status: OK

Figure 5. Classification results of MRMR –Modified GA applied Data set.

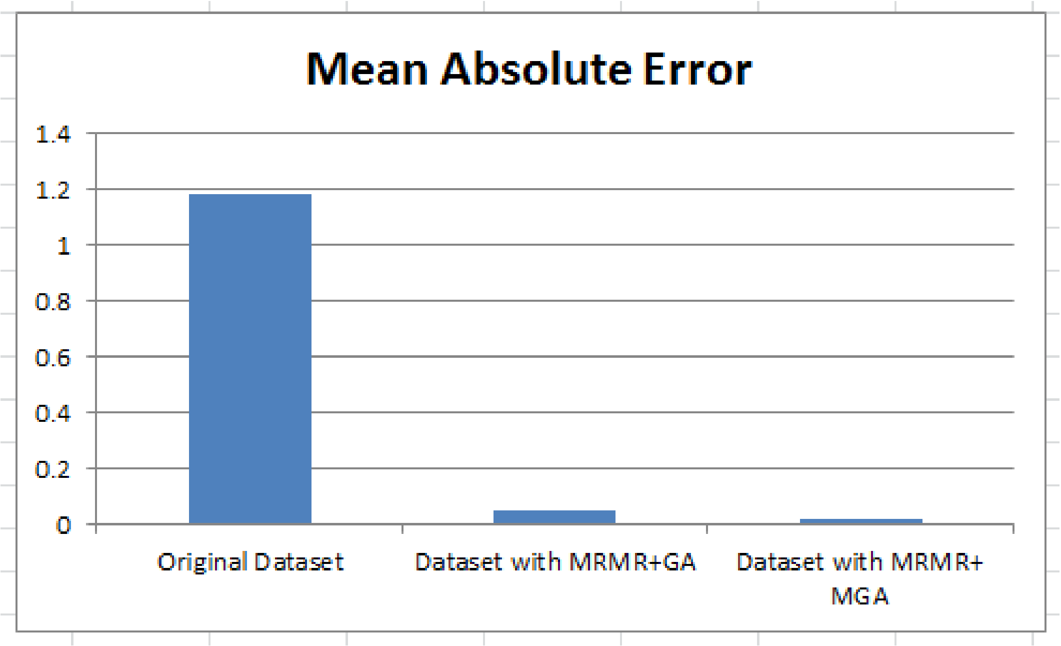


Figure 6. Error Rate of the Classifier.

The experimental results show that the accuracy of microarray data classification which has feature selection is better than without feature selection. MRMR with modified GA gives more accurate results. The error rate of the classification is also reduced using the proposed method.

Conclusion

This paper implements the proposed MRMR with modified GA as a feature selection method for microarray classification. Then, a Random Forest is used to assess the classification performance. Experimental results showed that the proposed method effectively simplifies gene selection and the total number of required parameters, thereby achieving higher classification

DOI: https://doi.org/10.52756/ijerr.2024.v39spl.006

accuracy compared to other feature selection methods. The classification accuracy obtained by the proposed method was higher than other methods. In the future, other available or modified filter and wrapper-based feature selections can be integrated and tested with other available or modified supervised classifiers.

Conflict of Interest

The authors declare no conflict of interest.

References

Albadr, M. A., Tiun, S., Ayob, M., & Al-Dhief, F. (2020). Genetic algorithm based on natural selection theory for optimization problems. *Symmetry*, *12*(11), 1758. https://doi.org/10.3390/sym12111758

Almugren, N., & Alshamlan, H. (2019). A survey on hybrid feature selection methods in microarray gene expression data for cancer classification. *IEEE Access*, *7*, 78533-78548. https://doi.org/10.1109/ACCESS.2019.2922987

Alromema, N., Syed, A. H., & Khan, T. (2023). A hybrid machine learning approach to screen optimal predictors for the classification of primary breast tumors from gene expression microarray data. *Diagnostics*, *13*(4), 708. https://doi.org/10.3390/diagnostics13040708 Balcha, A. and Woldie, S. (2023). Impact of Genetic Algorithm for the Diagnosis of Breast Cancer: Literature Review. *Advances in Infectious Diseases*, *13*, 41-46. https://doi.org/10.4236/aid.2023.131005.

Bhartiya, R., & Prajapati, G. L. (2023). NNFSRR: Nearest Neighbor Feature Selection and Redundancy Removal Method for Nearest Neighbor Search in Microarray Gene Expression Data. *EAI Endorsed Transactions on Pervasive Health and Technology*, *9*(1). http://dx.doi.org/10.4108/eetpht.9.3910

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324

Brown, E. C., & Sumichrast, R. T. (2005). Evaluating performance advantages of grouping genetic algorithms.*Engineering Applications of Artificial Intelligence*, *18*(1), 1-12. https://doi.org/10.1016/j.engappai.2004.08.024

Cerrada, M., Zurita, G., Cabrera, D., Sánchez, R. V., Artés, M., & Li, C. (2016). Fault diagnosis in spur gears based on genetic algorithm and random forest. *Mechanical Systems and Signal Processing*, *70*, 87-103. https://doi.org/10.1016/j.ymssp.2015.08.030

Cheng, J. H., Sun, D. W., & Pu, H. (2016). Combining the genetic algorithm and successive projection algorithm for the selection of feature wavelengths to evaluate exudative characteristics in frozen–thawed fish muscle. *Food Chemistry*, *197*, 855-863. https://doi.org/10.1016/j.foodchem.2015.11.019

El Akadi, A., Amine, A., El Ouardighi, A., &Aboutajdine, D. (2011). A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. *Knowledge and Information Systems*, *26*, 487-500. https://doi.org/10.1007/s10115-010-0288-x

Ghaheri, A., Shoar, S., Naderan, M., & Hoseini, S. S. (2015). The Applications of Genetic Algorithms in Medicine. *Oman Medical Journal*, *30*(6), 406–416. https://doi.org/10.5001/omj.2015.82

Hajieskandar, A., Mohammadzadeh, J., Khalilian, M., & Najafi, A. (2020). Molecular cancer classification method on microarrays gene expression data using hybrid deep neural network and grey wolf algorithm. *Journal of Ambient Intelligence & Humanized Computing*, *14*(5), 5297–5307. https://doi.org/10.1007/s12652-020-02478-x

Hameed, S.S., Hassan, R., Hassan, W. H., Muhammadsharif, F. F., & Latiff, L. A. (2021). The microarray dataset of ovarian cancer in csv format. *PLOS ONE*. Dataset. https://doi.org/10.1371/journal.pone.0246039.s006

Li, X., & Yin, M. (2013). Multiobjective binary biogeography based optimization for feature selection using gene expression data. *IEEE Transactions on NanoBioscience*, *12*(4), 343-353. https://doi.org/10.1109/tnb.2013.2294716

Liu, S., Tai, H., Ding, Q., Li, D., Xu, L., & Wei, Y. (2013). A hybrid approach of support vector regression with genetic algorithm optimization for aquaculture water quality prediction.*Mathematical and Computer Modelling*, *58*(3-4), 458-465. https://doi.org/10.1016/j.mcm.2011.11.021

Liu, X. Y., Liang, Y., Wang, S., Yang, Z. Y., & Ye, H. S. (2018).A hybrid genetic algorithm with wrapper-embedded approaches for feature selection. *IEEE Access*, *6*, 22863-22874. https://doi.org/10.1109/ACCESS.2018.2818682

Mandal, M., & Mukhopadhyay, A. (2013).An improved minimum redundancy maximum relevance approach for feature selection in gene expression data.*Procedia Technology*, *10*, 20-27. https://doi.org/10.1016/j.protcy.2013.12.332

Mishra, V., Mishra, M., Sheetlani, J., Kumar, A., Pachouri, P., Nagapraveena, T., Puttamallaiah, A., Sravya, M., & Parijatha, K. (2023). The Classification and Segmentation of Pneumonia using Deep Learning Algorithms: A Comparative Study. *Int. J. Exp. Res. Rev.*, *36*, 76-88. https://doi.org/10.52756/ijerr.2023.v36.007

Osama, S., Shaban, H., & Ali, A. A. (2023). Gene reduction and machine learning algorithms for cancer classification based on microarray gene expression data: A comprehensive review. Expert Systems with Applications, 213, 118946.
https://doi.org/10.1016/j.eswa.2022.118946

Shukla, A. K., Singh, P., &Vardhan, M. (2020). Gene selection for cancer types classification using novel hybrid metaheuristics approach. *Swarm and Evolutionary Computation*, *54*, 100661. https://doi.org/10.1016/j.swevo.2020.100661

Syahidin, Y., Maulidevi, N. U., & Surendro, K. (2023). Feature selection method based on genetic algorithm with wrapper-embedded technique for medical record classification. *In Proceedings of the 2023 12th International Conference on Software and Computer Applications*, pp. 184-191.
https://doi.org/10.1145/3587828.3587856

Tyagi, K., Kumar, D., & Gupta, R. (2024). Application of Genetic Algorithms for Medical Diagnosis of Diabetes Mellitus. *International Journal of Experimental Research and Review, 37(Special Vol)*, 1-10. https://doi.org/10.52756/ijerr.2024.v37spl.001

Welikala, R., Fraz, M., Dehmeshki, J., Hoppe, A., Tah, V., Mann, S., Williamson, T., & Barman, S. (2015). Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy. *Computerized Medical Imaging and Graphics*, *43*, 64–77. https://doi.org/10.1016/j.compmedimag.2015.03.003

Zare, M., Azizizadeh, N., & Kazemipour, A. (2023). Supervised feature selection on gene expression microarray datasets using manifold learning. *Chemometrics and Intelligent Laboratory Systems*, *237*, 104828.
https://doi.org/10.1016/j.chemolab.2023.104828

Ziegler, A., & König, I. R. (2014). Mining data with random forests: current options for real- world applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(1), 55-63. https://doi.org/10.1002/widm.1114

How to cite this Article:

P. Nancy Vincentina Mary and R. Nagarajan (2024). Feature Selection in Microarray using Proposed Hybrid Minimum Redundancy-Maximum Relevance (MRMR) and Modified Genetic Algorithm (MGA). *International Journal of Experimental Research and Review*, *39*(spl.) 82-91.

DOI : https://doi.org/10.52756/ijerr.2024.v39spl.006