MICROARRAY GENE CLASSIFICATION FOR A HYBRID ALGORITHM

Narayan Naik[1]*, Sharath Kumar Y H[2]

Abstract

In microarray gene expression analysis, a challenging issue has always been the feature's high dimensionality with a restricted sample size. For microarray datasets to be accurately classified, a reliable and effective feature selection method must be created. The maximum relevance (mRMR), minimum redundancy and adaptive Genetic Algorithm (AGA) are used in the hybrid feature selection technique known as mRMRAGA. The technique known as mRMR is widely used to more precisely determine the phenotypic traits of genes. The method by which feature relevance is reduced and described when paired with their pertinent feature selection is known as the maximum relative margin of rejection. Natural selection, which relies on heuristic search techniques, served as the model for the Genetic Algorithm (GA). The Adaptive genetic algorithms are genetic algorithms that have been modified and applied in the part that follows.

In this paper, the experiment was carried out using four benchmarked microarray gene expression datasets. One of these datasets has two class labels, while the other three have more than two. This indicates that the number of class labels in these datasets is heterogeneous.

[1]*Assistant Professor, Department of IS&E, Canara Engineering College, Bantwal-574219 Visvesvaraya Technological University, Belagavi-590018, Karnataka, India. Email: naik.mtech09@gmail.com

[2]Professor, Department of ISE, MIT Mysore-574177. Visvesvaraya Technological University, Belagavi-590018, Karnataka, India. Email: sharathyhk@gmail.com

***Corresponding Authors:** Narayan Naik

*Assistant Professor, Department of IS&E, Canara Engineering College, Bantwal-574219 Visvesvaraya Technological University, Belagavi-590018, Karnataka, India. Email: naik.mtech09@gmail.com

DOI: 10.53555/ecb/2022.11.12.252

Introduction

As a basis for prediction and diagnosing cancer, the microarray method is considered the gold standard in bioinformatics. Many times, the cancer's diagnosis and prognosis have been contrasted to the categorization of microarray gene expression datasets (Heller 2002, Li and Li 2008). The microarray dataset encompasses information pertaining to the levels of gene expression. Through data analysis, the location of altered genes can be found. The augmentation of classification as well as diagnostic methodologies is poised to unquestionably contribute substantial value to the domain of medical science for detecting disorders associated to certain genes, even as a biologist conducts an economical and effective examination of the gene expression levels of a small set of determined genes (Cosma et al. 2017). Classification and tumor type prediction, however, continue to be extremely difficult problems for medical science. For this purpose, determining the data profile of microarray gene expression is essential. However, numerous statistical techniques fell short in identifying a gene expression dataset's subset involved in disease. Because there were few samples compared to many genes (features). In the end, this makes microarray data analysis more difficult (Singh and Siva Balakrishnan 2015). Microarray datasets include redundant information and irrelevant genes/features, but this adds a significant amount of computational complexity (Wang 2012).

Given that other features already make these redundant data available, they actually don't help build a better predictor. (Song et al. 2011). Figure 5.1 displays a N × M matrix example of a microarray dataset. Here, the variable N denotes the samples number, M represents genes' number, and class labels denoted as {li | i = 1, 2, ⋯, N}.

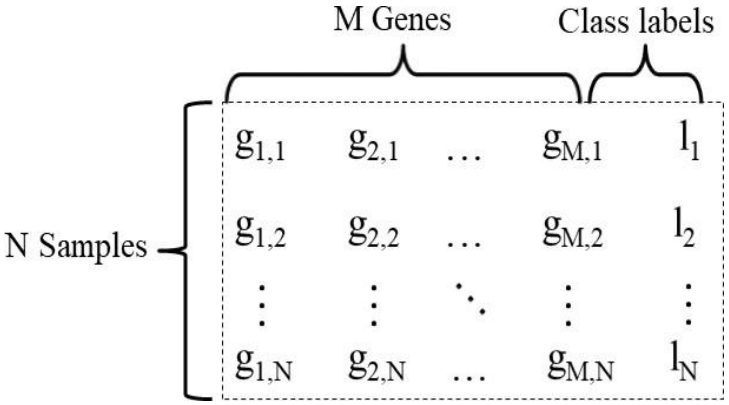

Fig. 5.1 $N \times M$ Microarray Dataset Example

The classification model's efficacy is frequently compromised by the deleterious influence of superfluous features inherent in datasets. Thus, a feature selection approach must be used to reduce features in order to improve a model's performance. Finding the important features subset is the main goal of feature selection, which has been identified as a highly important area of focus in bioinformatics and machine learning (Liu et al. 2018).

As stated by Saeys and colleagues (2007), in the feature selecting area, three methods are in use. Based on the standards that's been applied to the learning algorithm, these strategies are divided into three categories: wrapper, filter and hybrid-based (Hira and Gillies 2015). Because the filter method does not use to build classifiers, the predictor's performance might not be as expected (Lazar et al. 2012).

The filter technique is further separated into non-parametric and parametric methods, according to Hameed et al. (2018). The similar sample distribution of different classes has been the focus of a parametric filter-based approach, such as Bayesian, chi-squared, and Analysis of Variance (ANOVA) (Saeys et al. 2007). The wrapper approaches are known to be classifier dependent, meaning that the same performance might not be obtained for all other classifiers (de Paula Canuto and Santana 2014, Hameed et al. 2018, Saeys et al. 2007, Lazar et al. 2012, and Xiong et al. 2001). Additionally, if the total performance was deemed to be below grade level and preprocessing was not utilized, there might be The method named Minimum-Redundancy-Maximum-Relevance (mRMR) had been created with microarray datasets in mind (Ding and Peng 2005).

This feature selection technique was identified as unique. The filter strategy used in this method attempts to choose those with significant predictive and uncorrelated characteristics. In this method, the feature subsets with a low correlation among themselves (redundancy) and high correlation to a

class (relevance) are chosen. By employing minimum-redundancy-maximum-relevance, features are organized in this manner. Redundancy features are evaluated simultaneously for continuous and discrete features using mutual information and Pearson correlation coefficient. F-statistic and Mutual information are used to calculate the significance of features for both continuous and discrete features jointly. (Hoque et al., 2014) presented a feature selecting technique according to mutual information, coupled with a no-dominant sorting algorithm, termed MIFS-ND. An optimization methodology "Non-dominated Sorting Genetic Algorithm-II" was utilized in this method to choose features as per the criteria named maximum-relevance-minimum-redundancy (Deb et al. 2000). In this case, redundancy and relevance sorted list was ranked using the domination count and dominated count, respectively. Next, From each of the groups, a single gene was selected so that these genes could discriminate together (Ghalwash et al. 2016). In order to avoid redundant feature selection, features must first be clustered according to correlation or domain knowledge (such as gene ontology or molecular function) before employing this method. The Genetic Algorithm (GA) is the name of the population-based stochastic optimization technique. The natural selection process using fundamental genetic principles serves as the basis for the GA (Jakobovic´ and Golub 1999). Genetic algorithms perform two operations, referred to as mutation and crossover. Pc represents the crossover probability, and the mutation probability by Pm. Two prevalent issues with GA are non-convergent and pre-mature convergence. These issues could arise because Pc and Pm's values might not have been set appropriately. The term "Adaptive Genetic Algorithm" refers to the process of modifying Pm and Pc values in order to enhance the traditional GA. Because of its capacity for adaptation, the AGA is more reliable and augments the finding of the globally perfect solution. Combining multiple tried-and-true algorithms to create a novel approach to solving challenging problems is known as the hybrid approach. With the advantages of conventional algorithms in mind, the hybrid algorithm was developed making it more reliable than traditional methods. This section has elucidated a hybridized methodology for feature selection, amalgamating the AGA with mRMR algorithm to obtain the lowest level of redundancy and the relevance of the maximum level in microarray datasets. By contrasting the classifier's accuracy with other refined techniques now in use, the proffered mRMRAGA feature selection technique's effectiveness has been demonstrated. To test the suggested approach, four benchmark datasets were subjected to three fine-tuned classification models.

System Architecture of Proposed Method

The architecture of the system's presented in figure 5.2. This subsystem receives input in microarray gene expression datasets form. To deal with the inconsistency and noise, the data were preprocessed and normalized. Subsequently, Techniques for feature selection that rely on correlation coefficients were employed to identify the features' (genes') dependency. Following that, the mRMRmethod was used as a feature selection for determining the importance of features, and subsequently to optimize the outcomes the algorithm AGA was utilized. Finally, efficiency and effectiveness were measures for different classifiers.

The proposed system was broadly divided into three parts

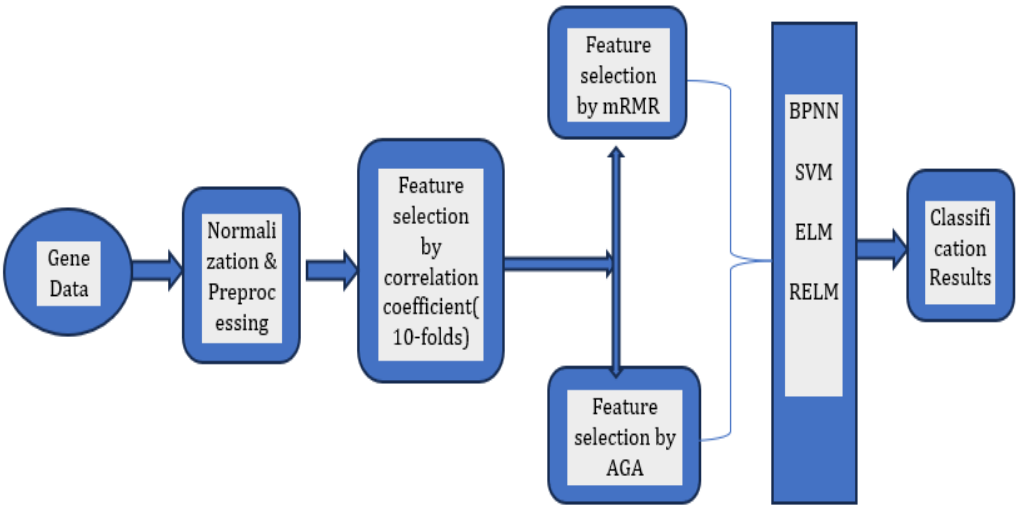

Fig. 5.2 System Architecture of Proposed Method

- Feature selection
- AGA Optimization
- Classification

The sections following will provide a thorough explanation.

Mutual Information as Computational metric of Relevance and Redundancy

Relevant features' identification is a key aspect of microarray gene expression probe. Identifying the most pertinent features was a crucial step in gene expression data analysis. The identification of genes (features) concerning class labels constitutes as primary objective of the method for choosing features, which produce the greatest amount of information. It has been discovered that feature entropy is an appropriate metric for identifying these genes (features). According to Cover and Thomas (1991), the entropy is simply the target class's initial uncertainty measure. The following equation 5.1 elaborates upon the concept of entropy:

$$H(X) = \sum_{x=1}^{N_x} P_x(X) \log(P_x(X))$$

For example, Px(x)jx = 1; 2; •••; Nx, where class probability. Equation 5.2 describes the feature vector, and for the vector, the average variance is used to compute the conditional probability as follows:

$$H(S|X) = \sum_{s=1}^{N_s} P(s) \left(\sum_{x=1}^{N_x} P_x(x|s) log(P_x(x|s)) \right)$$

The is Px(x|s) of class x is the conditional probability, and there are Ns samples in input feature vector, which is 's'. Typically, the initial entropy is anticipated to surpass the conditional entropy. As long as the conditional entropy value equals the original entropy value, the features are considered independent of the classes. Accordingly, mutual information determines the amount of reduced uncertainty (Battiti 1994). Equation 5.3 yields the variables x to s mutual information I(X; S) as follows:

$$I(X; S) = H(X) – H(X|S) \quad (5.3)$$

The rewording of Equation 5.3 is:

$$I(X; S) = I(S; X) = \sum P(x, s) \log \frac{P(x, s)}{P(x)P(s)} \quad (5.4)$$

With regard to S and X, the mutual information function's symmetry property, I(S; X) = I(X; S).

Minimum-redundancy-maximum-relevance (mRMR)

The popular feature selection process known as mRMR (minimum relevancy and minimum redundancy) with mutual information quotient (MIQ) and mutual information difference (MID) is described in this section (Paul and Iba 2005). According to Li et al. (2007), genes that exhibit notable variations in expression across two distinct classes—tumor and normal, or cancer subtypes—are referred to as deferentially expressed genes. The degree of deferential expression of a gene is interpreted as an indicator of its relevance. By calculating the mutual information, a gene's significance can be ascertained (Ding and Peng 2005, Thomas and Cover 1991). Regarding other classes, the gene's mutual information is zero if it is distributed evenly or uniformly across them. In this case, only the discrete variable has been considered when considering the mutual information. X and S are two discrete variable characteristics, and equation 5.4 specifies their mutual information I. The method by which features (genes) are chosen so that their mutual information has the greatest dissimilarity degree concerning other genes is known as the principle of minimizing redundancy. The subset of genes that need to be identified is shown here as s. Equation 5.5 gives the average minimal redundancy.

$$minimum(W) = \frac{1}{|s|^2} \sum_{i,j \in s} I(i, j)$$

Whereas |s| represented the genes in S and I(i, j) illustrated how the ith and jth genes mutual information between them. Once again, the deferentially expressed gene can be chosen using the mutual information. I(h, gi), the mutual information, is described by equation 5.6, which was identified as the discriminant control of genes. The gene relevance is computed using the mutual information between gene expression gi and target classes h1, h2 ... hk. Therefore, average relevance maximization is equivalent to the relevance maximization for every gene in subset, that's been determined by 5.6 equation as follows:

Therefore, the urgent need for improved classification accuracy is to minimize gene redundancy and maximize gene relevance. This indicates that although both criteria are equally important, they are not distinct from one another; rather, they may have been combined to form an individual parameter in mRMR. Consequently, two straight forward combined criteria are established by Max(V/M) and Max(V – W).

Currently, mRMR has been defined in terms of mRMRMID and mRMRMIQ for discrete data. Equations 5.7 and 5.7, respectively, provide the formulas for mutual information quotient and mutual information difference.

$$mRMR_{MID} = max_{i\in\Omega_s}\left[I(i,h) - \frac{1}{|s|}\sum_{j\in s} I(i,j)\right]$$

$$mRMR_{MIQ} = max_{i\in\Omega_s}\left\{I(i,h)/\left[\frac{1}{|s|}\sum_{j\in s} I(i,j)\right]\right\}$$

The mRMR Algorithm is defined as illustrated in Algorithm 4:

Algorithm 4: Feature selection with mRMR

Input: n - number of features to be selected, d - discretized data, g - number offeatures in d, c - class
Output: F - feature set.
idleft = [1: *g*];
for (*i* = 1: *g*): **do**
relevance(*i*) = *mutualinfo* (*d*(:,*i*), *c*) ;
end
[*R, id*] = *Max*(*relevance*);*F* [1] = *id*;
idleft = *idleft* − *F*;
for (*i* = 2: *n*): **do**
obj_1 = *relevance*(*idleft*);
for (*j* = 1: |*idleft*|): **do**
sum = Σ_F (*mutualinfo*(*d*(:, *k*), *d*(:, *idleft*)) ;
redun(*j*) = *sum*/|*F* |;
end
obj_2 = *relevance*(*idleft*)/(*redun* + 0.0001) ;
[*newid*, obj_2] = *Nondominated Feature Selection* (obj_1, obj_2, *idleft*);
[*R, id*] = *Max*(obj_2);
F [*i*] = *id*;
idleft = *idleft* − *F*;
end

Adaptive Genetic Algorithm (AGA)

The two main functions of a genetic algorithm (GA) are mutation and crossover. These operations mutation and crossover, respectively, generate new individuals both globally and locally. These two actions ensured that local and global searches would be conducted for the GA. The mutation probability (Pm), as well as crossover probability (Pc), were utilized to determine whether the genetic algorithm had converged to identify the optimal solution. In a standard Genetic Algorithm, Pc and Pm are kept constant parameters in Genetic Algorithm the searching procedure. Pc will make the global search too coarse when it gets extremely large, which can make it nearly impossible to get the optimum result. Furthermore, searching within local minima may be lost if Pc is extremely small. On the other hand, the Genetic Algorithm acts like a random search when Pm is extremely high. Furthermore, the exploratory capacity of searching will be inhibited if Pm is extremely small.

Multiple cross-validations are needed to find the best values for Pc and Pm. Adaptive genetic algorithms (AGAs) are a more appropriate process when the Genetic Algorithm is given the freedom to modify Pc and Pm values throughout the search space. Equations 5.9 and 5.10 provide a formula for adjusting the values of Pc and Pm.

$$P_c = \begin{cases} K_1\frac{(f_{max}-f')}{(f_{max}-f_{avg})}, & f' \geq f_{avg} \\ K_2, & f' < f_{avg} \end{cases}$$

$$P_m = \begin{cases} K_3\frac{(f_{max}-f)}{(f_{max}-f_{avg})}, & f \geq f_{avg} \\ K_4, & f < f_{avg} \end{cases}$$

Where, in chromosome crossover, the greater parents' fitness is referred to as f j, fmax stands for the maximal fitness of every individual, and "favg" stands for "average fitness." (Montana and Davis 1989). The four control variables, denoted as k1, k2, k3, and k4, have a range of 0 to 1. Figure 5.3 shows the AGA optimization procedure.

Hybrid Feature selection (mRMRAGA)

This chapter proposed mRMRAGA selection, a hybrid gene selection strategy that combined e best

features of AGA and mRMR. The classifier of choice is the ELM, and the ELM's classification accuracy is determined by calculating its AGA fitness. In formulas 5.9 and 5.10, K1 value 0.9, K2 value 0.6, K3 value 0.1, and K4 value 0.001. There can be up to 600 iterations in total. Assuming that a1 and a2 are samples from the gene's expression dataset A. The ensuing elucidation provides an intricate account of the mRMRAGA selection process:

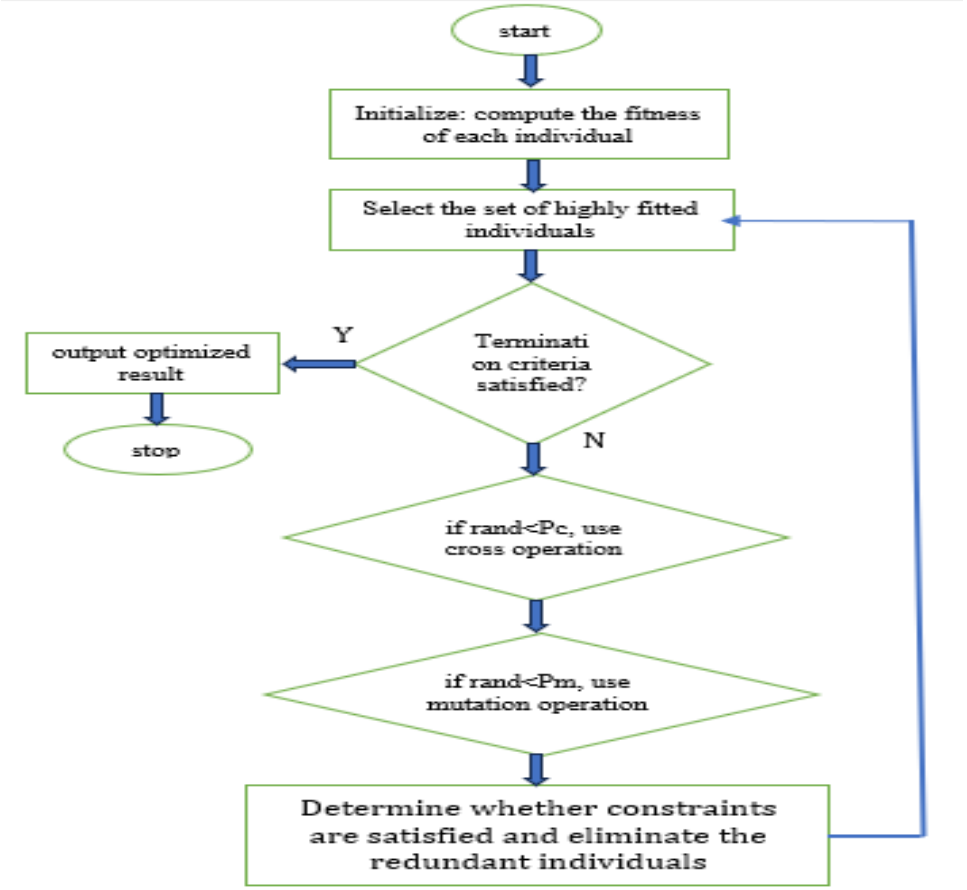

Fig. 5.3 AGA Optimization Process

1. For every gene in data set A, mutual information is calculated. Repeatedly applying mRMR yields the subset B of A. Assume 300 as B feature number.
2. After initializing the AGA population, the fitness is calculated for every individual. Have determined the population size based on problem space. Attaining the optimal solution may be facilitated when dealing with a problem of considerable magnitude, but it will take some time to complete the AGA search. This population has a size of 30. A gene's sample size is equal to a1, and every individual possesses numerous 'B' features.
3. Binary coding is used to encode a population of thirty (30) individuals. Following that, every single one looks similar as a chromosome having 300 length.
4. For 'fmax', 'favg', and 'f', the fitness values are calculated.
5. Curate a subset of exceptionally qualified entities through the establishment of a predefined threshold.
6. The individuals are randomly paired in step (5), and a new population is created by using crossover operations based on the Pc value.
7. A mutation operation is applied based on the Pm value to create a new population.
8. Verify that the termination criterion or current optimal fitness value is satisfied. Proceed to step (9), if true; if not, proceed to step (4).
9. The ideal subsets of genes are determined by the decoding principles.

Experimental Evaluation

This section furnishes the prescribed methodologies for the experimental validation. In the context of the experiments, the 4 extensively utilized benchmark datasets were chosen. The class numbers in these datasets are heterogeneous, meaning that while some have two class labels, others have more than two. The section below contains descriptions of the dataset. The sections that follow go into more detail about the experimental settings.

Datasets

For the purpose of conducting comprehensive experiments, four benchmarked gene expression microarray datasets were chosen. Scientists working in this area frequently use these datasets. Researchers have public access to these datasets. Table 5.1 describes these cancer datasets, which are Lymphoma, Lung and SRBCT.

Zhu et al. (2007a) provided datasets on lung and breast cancer, which were utilized in this investigation. The breast cancer dataset incorporates samples from 97 patients and comprises 24,481 features, with two distinct class labels: 51 for cancer cases and 46 for normal cases. 181 patient samples are included in lung cancer datasets. Class labels (20-COID, 139-AD, 6-SMCL, 17-NL, and 21-SQ) are present and 12533 features in this dataset. Where AD stands for adenocarcinoma, SQ for squamous cell carcinoma, 'NL' normal lung, COID for pulmonary carcinoma, and SMCL for small cell lung cancer. The microarray dataset for lymphoma was sourced from Dortling and Buhlmann (2002) and utilized in their investigation. There are 62 samples and 4026 features in this dataset. The three distinct adult lymphoid malignancies represented by these samples are 9-FL, 42-DLBCL, and 11-CLL. Where DLBCL stands for diffuse large B-cell lymphoma, FL for follicular lymphoma and CLL for chronic lymphocytic leukemia. The SRBCT dataset utilized in the study was sourced from D´ıaz-Uriarte and De Andres (2006). There are 63 samples and 2308 features in this dataset. Four classes comprise these samples: 20 RMS, 23 EWS, 12 NB and 8 NHL.

Table 5.1 The four dataset's characteristics

Datasets	Number of Features	Number of Samples	Number of Classes	Class Description	References
Breast	24481	97	2 (46- 51)	46 Normal 51 Cancer	Zhu et al. (2007*a*)
Lung	12600	203	5 (139-17-6-21-20)	139 AD 17 NL 6 SMCL 21 SQ 20 COID	Zhu et al. (2007*a*)
Lymphoma	4026	62	3 (42-9-11)	42 DLBCL 9 FL 11 CLL	Dettling and Bühlmann (2002)
SRBCT	2308	63	4 (23-8-12-20)	23 EWS 8 NHL 12 NB 20 RMS	D´ıaz-Uriarte and De Andres (2006)

Data Preprocessing and Feature selection

The complete methodology that has been carried out in this chapter is shown in figure. The experimental protocol is described in the section below.

- The datasets underwent 10 cycles of preprocessing using the Pearson Correlation Coefficient approach as the first step in the data analysis procedure. First and foremost, to ensure that every dataset traverses this stage, which will result in a precision output at the reduction end. Various thresholds have been investigated for the quantity of chosen genes. To test the accuracy of the performance, varying numbers of genes were filtered in each iteration for every dataset.
- The reduced datasets have been assessed over the applied classifiers to compare the performance on different parameters.
- The correlation coefficient, a different feature selection method, was used to further refine these condensed datasets. Several classifiers were used to derive the fitness function for the genetic algorithm. 10-folds cross-validation was carried out to guarantee that training and testing make use of the complete datasets.
- The suggested method, mRMRAGA, which combines mRMR and AGA, further reduced the features
- Lastly, a comparison was made between the various classifiers' classification outcomes.

Classifiers

The experiments were carried out using the four fine-tuned classifiers. Since no single technique is effective on all datasets, a variety of classifiers have been chosen for the tests, and no single classifier performs consistently across all datasets. (SVM) Support vector machines, BPNN (backpropagation neural networks), ELM (extreme learning machines) and regularized extreme learning machines (RELM) are the classifiers that are being used in this instance. These working principles of classifiers have all been explained.

Artificial Neural Network (ANN)

Another example of a linear model derived from natural neurons is ANN. An artificial neuron, or perceptron, is a group of interconnected perceptron's that make up an ANN. An artificial neural network's output is the weighted sum of its

perceptron connections. The term "hidden layer" refers to the group of perceptron's that connects the input nodes to the output nodes. Figure 5.4 depicts a simple ANN with a single hidden layer. Finding the best set of weights for a given situation and training the ANN, the backpropagation technique is frequently employed (Cilimkovic 2015). Deep learning is the term used to describe an ANN with multiple hidden layers. It is this area of machine learning that is currently undergoing the most research.

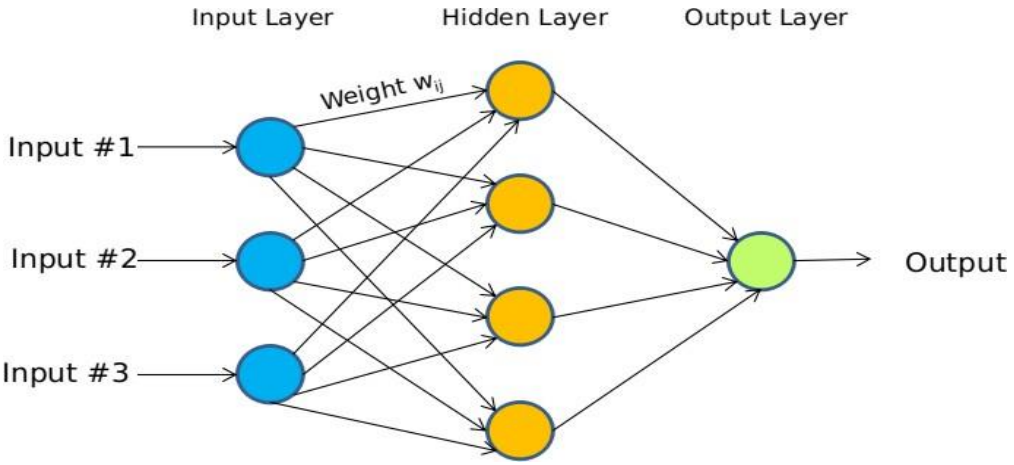

Fig. 5.4 Artificial Neural Network

Back Propagation Neural Network (BPNN)

Back-propagation is the fundamental technique used in neural net training. A neural network's weights are adjusted according to the error that was obtained in the preceding iteration (epoch). Proper tuning of weights lowers the error rates and improves the model's reliability by improvising its generalization. Figure 5.5 displays the schematic diagram for the BPNN.

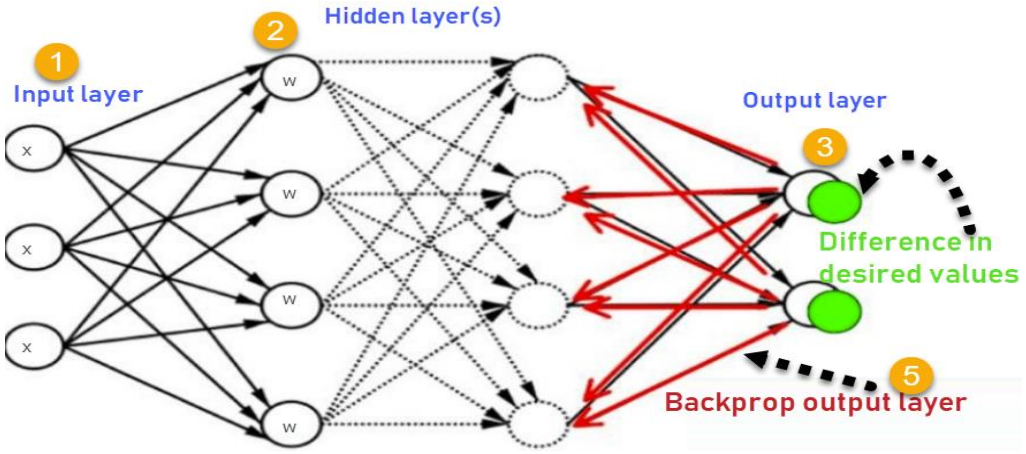

Fig. 5.5 Back Propagation Neural Network

Backpropagation is the abbreviation for "backward propagation of errors." Backpropagation is the standard training technique for artificial neural networks. Every network weight is calculated to determine the gradient of the loss function. The BPNN approach is susceptible to data noise.

Extreme Learning Machines (ELM)

ELM are the emergence of a significant machine learning approach. The model's parameter can be calculated without the learning process. The primary benefit of these methods is this. EML is actually the abbreviation for a Single-Layer Feed-Forward Neural Network (SLFN). The fundamental tenet of ELM is that training data is independent since the hidden layer's weight doesn't need to be adjusted. As long as there are sufficient training data and hidden neurons to identify each hidden neuron's parameter, ELM can accurately solve any regression problem. The universal approximation property is what is meant by this. In figure 5.6, the basic neural network is displayed.

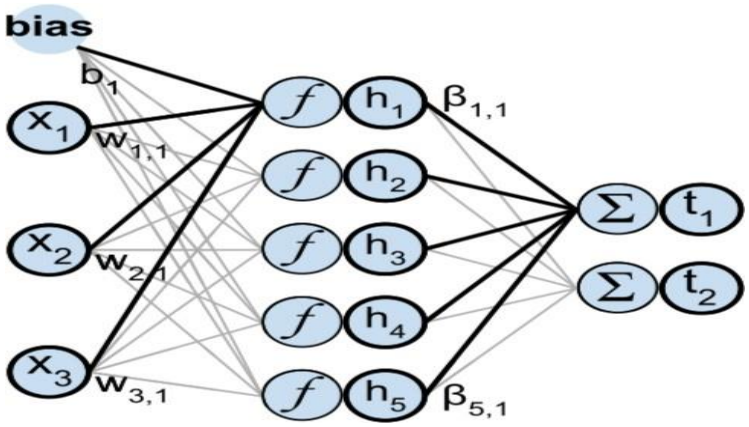

Fig. 5.6 Neural Networks

Additional benefits of EML include decreased negative effects from overfitting, random initialization, and regularization of the model structure. When N training samples (x, t) is used, equation 5.11 shows the SLFN outputs having L hidden neurons.

$$\sum_{j=1}^{L} \beta_j \phi(w_j x_i + b_j), \quad i \in [1, N]$$

The relationship between the target and the network's inputs and outputs is shown in Equation 5.12: The data input is transformed into a different representation by the hidden neurons in two steps. A hidden layer receives the data through the input layer's biases and weights, and the outcome is then subjected to the non-linear activation function. In an experiment, ELMs with a matrix form of the equation have been analyzed as regular neural networks. The matrix form is represented by equations 5.13 and 5.14. Figure 5.7 displays the schematic diagram of the ELM.

$$y_i = \sum_{j=1}^{L} \beta_j \phi(w_j x_i + b_j) = t_i + \epsilon_i, \quad i \in [1, N]$$

$$H = \begin{bmatrix} \phi(w_1 x_1 + b_1) \ldots . \phi(w_L x_1 + b_L) \\ \vdots \qquad \ddots \qquad \vdots \\ \phi(w_1 x_N + b_1) \ldots . \phi(w_L x_N + b_L) \end{bmatrix}$$

$$\beta = (\beta_1^T \ldots \beta_L^T)^T \ , \ T = (y_1^T \ldots y_L^T)^T$$

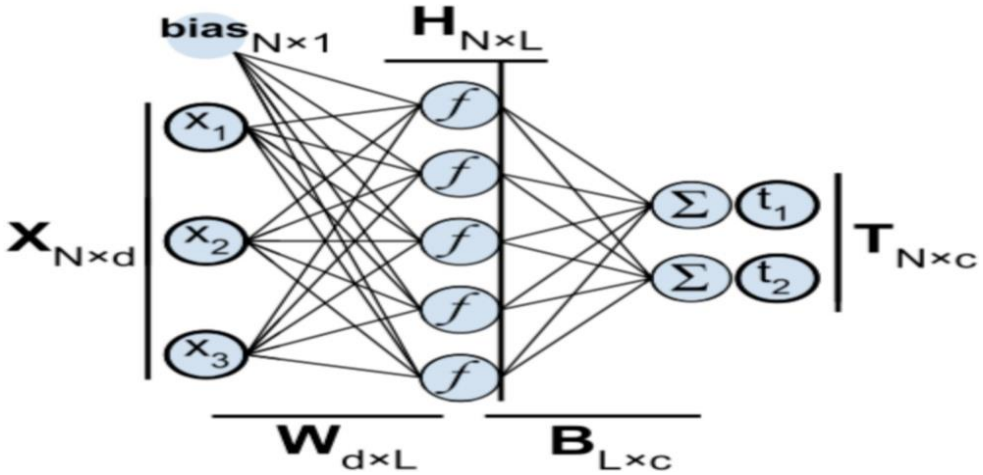

Fig. 5.7 Matrix of Extreme Learning Machines

Regularized ELM(RELM)

When encountering irrelevant or correlated data, ELM techniques might have some problems. A few of them were pruned by applying the L1 regularization to the hidden layer. OP-ELM (Optimally-Pruned ELM) is the common term used to describe this L1-regularized ELM. OPELM is the expanded version of the ELM algorithm, which incorporates neuronal pruning to strengthen the algorithm. (Rong et al. 2008) described a method for building an ELM-based network with pruning neurons for categorization applications. The relevance of the output was then assessed using statistical tests on the neurons.

Result and Discussion

On these four selected microarray datasets, Ten (10) runs of the mRMRAGA selection have been performed, each time, with a distinct target quantity of genes to be chosen. Table 5.2 contains a tabulation of the outcomes.

Table 5.2 The genes count chosen through mRMRAGA selection on microarray datasets.

Dataset	Number of Genes in step									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Breast	25	45	65	90	115	135	155	165	185	215
Lung	32	60	85	105	125	145	155	165	184	214
Lymphoma	10	35	60	85	110	130	150	170	186	208
SRBCT	20	45	70	95	120	140	160	180	195	210

Table 5.3 display the classification accuracy rates using the ELM classifier for each subgroup of the datasets. These classification accuracy metrics were determined by averaging the results of thirty repetitions of the classification process.

Table 5.3 Classification accuracy of mRMRAGA Selection and ELM

Dataset	Classification accuracy rates %									
	1	2	3	4	5	6	7	8	9	10
Breast	82.47	84.32	87.19	85.12	84.39	86.73	92.31	95.37	94.21	93.55
Lung	97.80	92.00	93.57	92.78	94.43	94.89	93.22	95.00	94.67	93.33
Lymphoma	95.34	94.80	94.00	92.88	92.00	91.50	90.00	89.32	88.50	88.24
SRBCT	94.66	95.80	90.11	89.09	86.36	87.16	88.07	88.98	88.64	88.52

To illustrate the mRMRAGA Selection algorithm's effectiveness, 3 properly calibrated feature selection algorithms were applied to identical datasets with similar target gene numbers. ReliefF, sequential forward selection (SFS), and mRMR are these feature selection techniques (Gu et al. 2014, Gu and Sheng 2016, Somol et al. 1999, Reunanen 2003). Tables 5.4, 5.5, and 5.6 below display the classification accuracy rate.

As a result, the mRMRAGA selection algorithm has a greater classification accuracy than other methods for choosing features like mRMR algorithms, ReliefF and SFS. The figure 5.8, 5.9, 5.10, and 5.11, correspondingly, compare the accuracy of the classification for the SRBCT cancer, breast, lymphoma and lung datasets.

Table 5.4 Classification accuracy of Relief and ELM

Dataset	Classification accuracy rates %									
	1	2	3	4	5	6	7	8	9	10
Breast	50.71	51.67	52.33	54.33	53.44	52.81	51.25	50.94	50.31	50.21
Lung	50.54	51.54	53.08	54.23	59.25	58.57	57.50	54.29	50.71	50.24
Lymphoma	52.34	53.12	55.30	56.00	57.40	56.70	55.32	54.38	52.86	51.76
SRBCT	58.32	59.87	68.04	62.51	65.39	64.24	63.44	60.39	59.63	58.61

Table 5.5 Classification accuracy of SFS and ELM

Dataset	Classification accuracy rates %									
	1	2	3	4	5	6	7	8	9	10
Breast	70.22	73.58	74.48	76.38	77.28	78.59	78.94	75.29	74.22	73.67
Lung	83.27	84.21	81.77	83.27	86.90	87.27	88.38	89.57	88.21	84.76
Lymphoma	82.65	84.32	83.21	85.12	84.73	83.45	84.00	86.88	85.35	85.12
SRBCT	81.48	86.77	85.28	86.68	82.08	79.27	80.26	83.43	80.33	79.12

Table 5.6 Classification accuracy of mRMR and ELM

Dataset	Classification accuracy rates %									
	1	2	3	4	5	6	7	8	9	10
Breast	80.00	70.59	73.56	72.31	75.65	73.21	76.33	77.89	73.43	72.89
Lung	79.52	77.94	77.22	77.14	78.33	77.22	78.50	78.61	75.62	74.61
Lymphoma	75.64	78.22	77.43	75.23	79.88	78.70	77.12	78.21	76.00	75.89
SRBCT	86.82	87.30	77.78	79.37	85.71	80.95	79.36	79.68	77.73	76.76

More evidence has been provided regarding the efficacy of the genes chosen using the mRMRAGA selection method. To categorize the chosen genes according to the recommended selection procedure (mRMRAGA), four well-tuned classification models were selected. These classifiers are Regularized Extreme Learning Machine (RELM), ELM, SVM, and BPNN. The accuracy of the classification for breast as well as cancer, lymphoma cancer, and SRBCT is illustrated in Figures 5.12, 5.13, 5.14, and 5.15, respectively.

The experiment's findings signifies that classification accuracy may not necessarily improve as the number of genes grows. An additional benefit is that the mapping of genes to classes is made easier when there are few genes present. The categorization rate is determined by the intricacy of the gene-to-gene association rather than complete genes selected. In the case of correlation identification, the accuracy curve is going to remain more steady, when the feature selection techniques closely match the classification model. Additionally, it has been noted that in this experiment, the RELM is a better classifier for the mRMRAGA selection strategy.

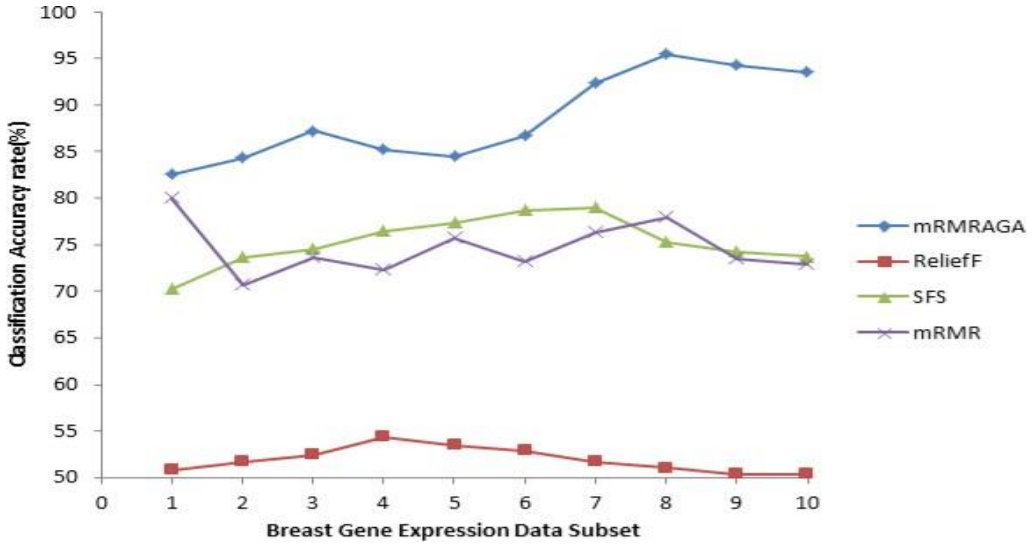

Fig. 5.8 Classification accuracy rates on the Breast cancer dataset with the use of feature selection algorithms

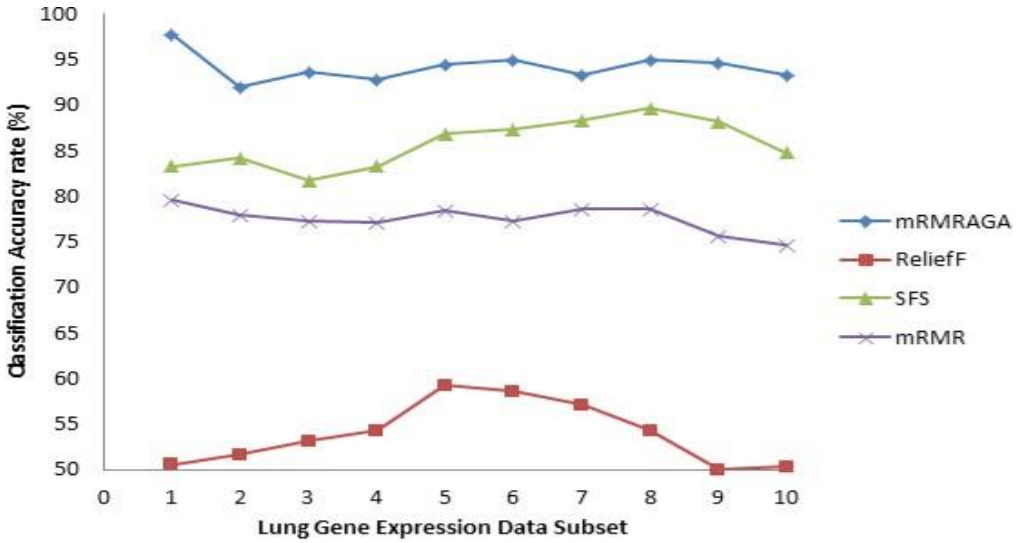

Fig. 5.9 Classification accuracy rates on the Lung cancer dataset with the use of feature selection algorithms.

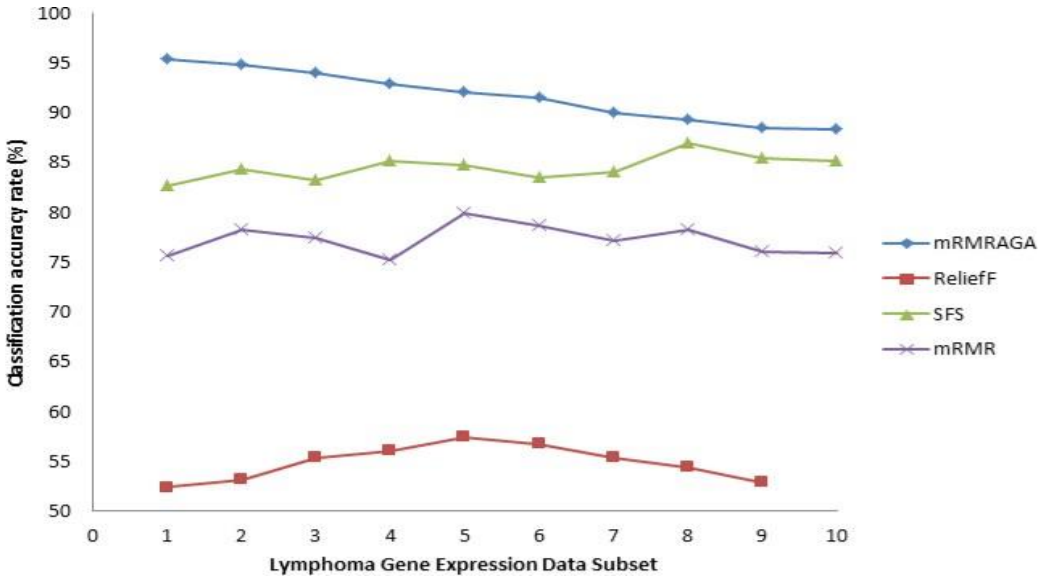

Fig. 5.10 Classification accuracy rates on the Lymphoma cancer dataset with the useof feature selection algorithms

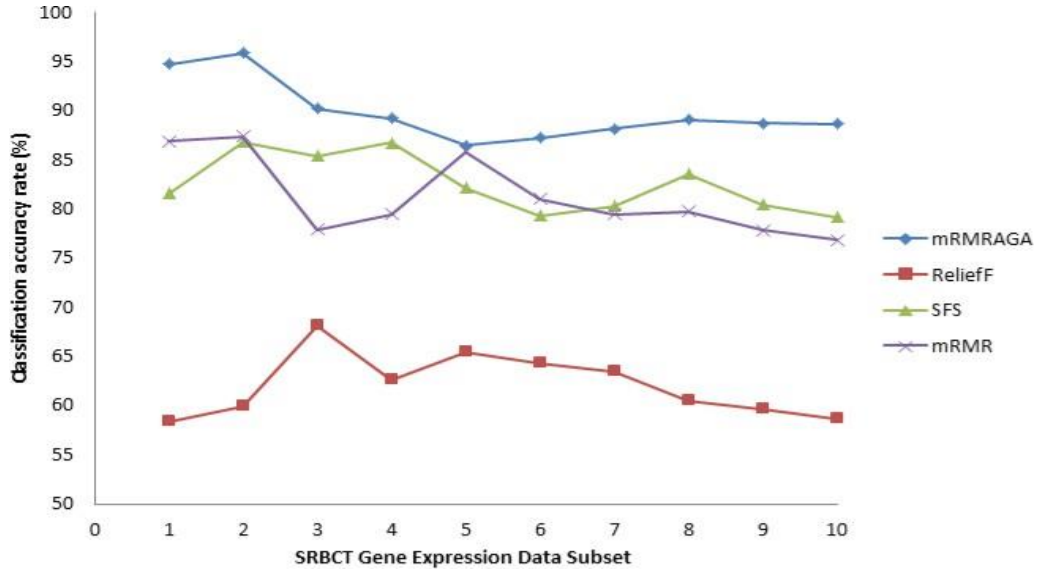

Fig. 5.11 Classification accuracy rates on the SRBCT dataset with the use of featureselection algorithms

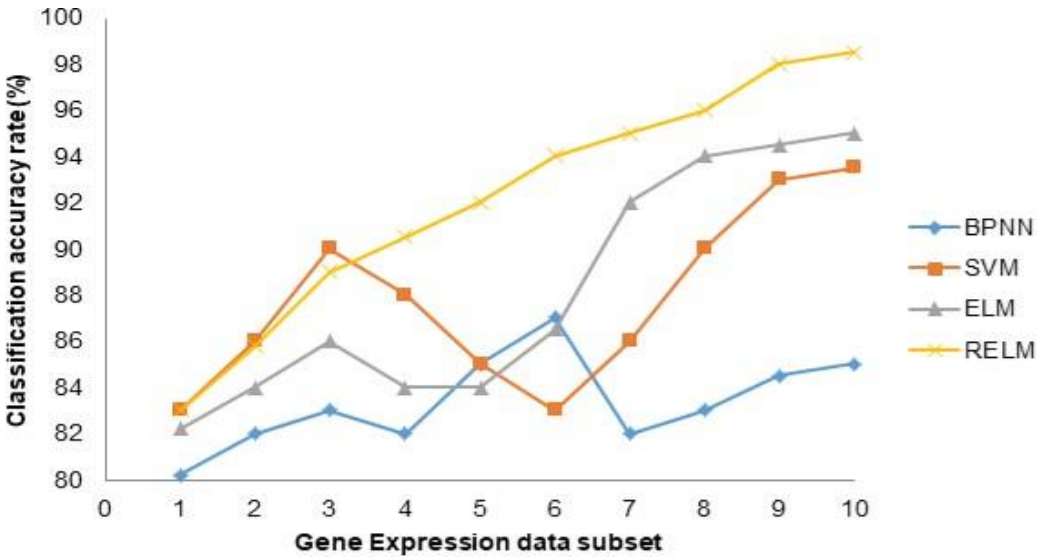

Fig. 5.12 Classification accuracy on Breast

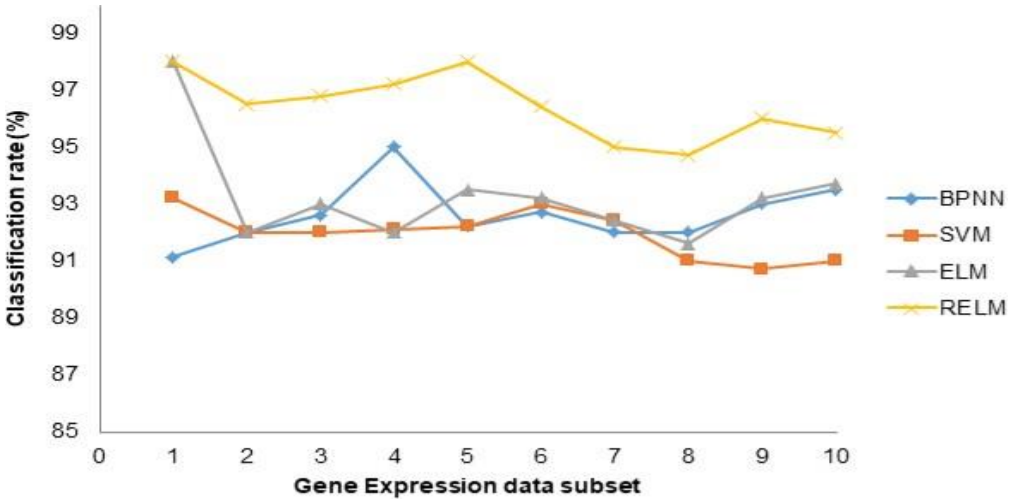

Fig. 5.13 Classification accuracy Lung

Summary

This section contains a synopsis of the research. Indeed, high-dimensional datasets, exemplified by gene expression databases, are characterized by genes that manifest a considerable count while being circumscribed by a restricted sample size. It therefore requires a unique and thorough analysis. The mRMRAGA method, a hybrid selection of features technique that blends mRMR and AGA, was proposed in this chapter. The mRMRAGA selection strategy successfully reduced the dimension while also lowering the dataset's redundancy, which increased the classification accuracy. For example,

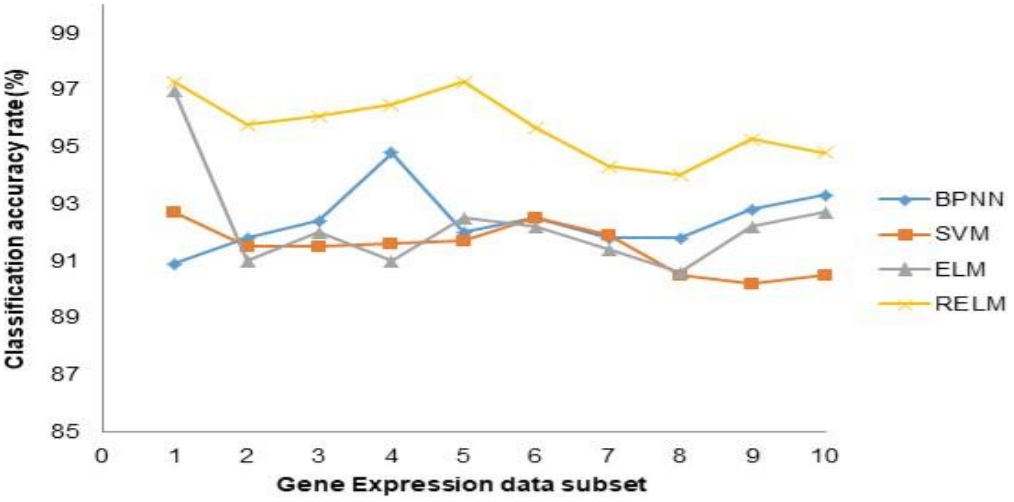

Fig. 5.14 Classification accuracy Lymphoma

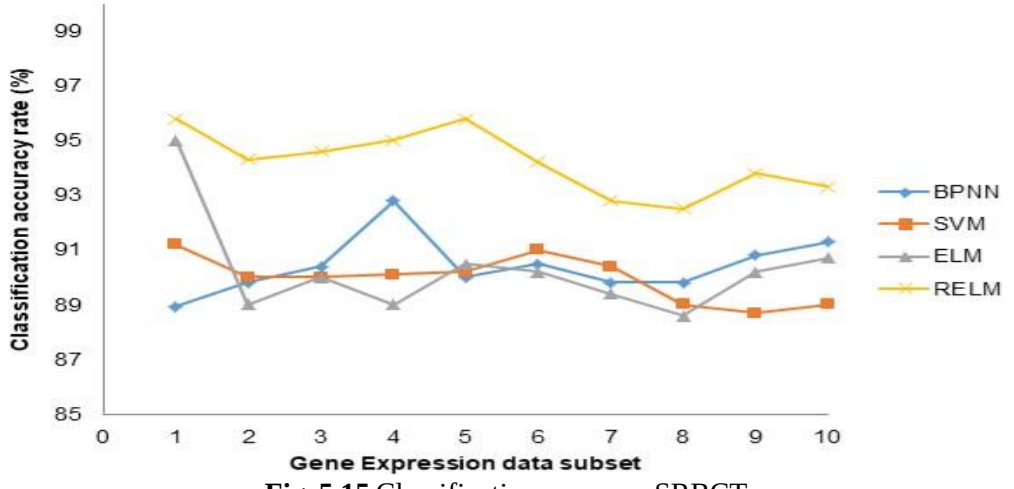

Fig. 5.15 Classification accuracy SRBCT

With a high degree of classification accuracy, the mRMRAGA selection approach can minimize the number of genes in a particular dataset that has about 20,000 genes to less than 300. Using four benchmarked datasets, the suggested feature selection method's efficacy was compared to the most advanced feature selection strategies, including ReliefF, SFS, and mRMR. Furthermore, it was noted that mRMRAGA outperformed in every instance. The primary feature selection process was then employed, and four distinct classifiers— RELM, BPNN, ELM, and SVM — have been employed. This indicates even more the stability of the proposed method.

References:

1. Heller, M. J. (2002), „Dna microarray technology: devices, systems, and applications", Annual review of biomedical engineering 4(1), 129–153.
2. Li, S. and Li, D. (2008), DNA microarray technology and data analysis in cancer research, World Scientific.
3. Cosma, G., Brown, D., Archer, M., Khan, M. and Pockley, A. G. (2017), „A survey on computational intelligence approaches for predictive modeling in prostate cancer", Expert systems with applications 70, 1–19.
4. Singh, R. K. and Siva Balakrishnan, M. (2015), „Feature selection of gene expression data for cancer classification: a review", Procedia Computer Science 50, 52–57.
5. Wang, L. (2012), Feature selection in bioinformatics, in „Independent Component Analyses, Compressive Sampling, Wavelets, Neural Net, Biosystems, and Nanoengineering X", Vol. 8401, International Society for Optics and Photonics, p. 840113
6. Song, Q., Ni, J. and Wang, G. (2011), „A fast clustering-based feature subset selection algorithm for high-dimensional data", IEEE transactions on knowledge and data engineering 25(1), 1–14.
7. Liu, S., Xu, C., Zhang, Y., Liu, J., Yu, B., Liu, X. and Dehmer, M. (2018), „Feature selection of gene expression data for cancer classification using double rbf-kernels", BMC bioinformatics 19(1), 396.
8. Saeys, Y., Inza, I. and Larranaga, P. (2007), „A review of feature selection techniques ˜ in bioinformatics", bioinformatics 23(19), 2507–2517.
9. Hira, Z. M. and Gillies, D. F. (2015), „A review of feature selection and feature extraction methods applied on microarray data", Advances in bioinformatics 2015.
10. Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C., de Schaetzen, V., Duque, R., Bersini, H. and Nowe, A. (2012), „A survey on filter techniques for feature selection in gene expression microarray analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 9(4), 1106–1119.
11. Hameed, S. S., Petinrin, O. O., Osman, A. and Hashi, F. S. (2018), „Filter-wrapper combination and embedded feature selection for gene expression data", Int. J. Advance Soft Compu. Appl 10(1).
12. Saeys, Y., Inza, I. and Larranaga, P. (2007), „A review of feature selection techniques ˜ in bioinformatics", bioinformatics 23(19), 2507–2517.
13. Xiong, M., Fang, X. and Zhao, J. (2001), „Biomarker identification by feature wrappers", Genome Research 11(11), 1878–1887.
14. Hameed, S. S., Petinrin, O. O., Osman, A. and Hashi, F. S. (2018), „Filter-wrapper combination and embedded feature selection for gene expression data", Int. J. Advance Soft Compu. Appl 10(1).
15. Xiong, M., Fang, X. and Zhao, J. (2001), „Biomarker identification by feature wrappers", Genome Research 11(11), 1878–1887.
16. Santana, L. E. A. d. S. and de Paula Canuto, A. M. (2014), „Filter-based optimization techniques for selection of feature subsets in ensemble systems", Expert Systems with Applications 41(4), 1622–1631.
17. Bolon-Canedo, V., S ´ anchez-Marono, N., Alonso-Betanzos, A., Ben ´ ´ıtez, J. M. and Herrera, F. (2014), „A review of microarray datasets and applied feature selection methods", Information Sciences 282, 111–135.
18. Ding, C. and Peng, H. (2005), „Minimum redundancy feature selection from microarray gene expression data", Journal of bioinformatics and computational biology 3(02), 185– 205. INTERNATIONAL JOURNAL OF SPECIAL EDUCATION Vol.37, No.3, 2022 - 14352 –
19. Hoque, N., Bhattacharyya, D. K. and Kalita, J. K. (2014), „Mifs-nd: A mutual information-based feature selection method", Expert Systems with Applications 41(14), 6371–6385.
20. Deb, K., Agrawal, S., Pratap, A. and Meyarivan, T. (2000), A fast elitist nondominated sorting genetic algorithm for multi-objective optimization: Nsga-ii, in „International conference on parallel problem solving from nature", Springer, pp. 849– 858.
21. Ghalwash, M. F., Cao, X. H., Stojkovic, I. and

Obradovic, Z. (2016), „Structured feature selection using coordinate descent optimization", BMC bioinformatics 17(1), 158.
22. Jakobovic, D. and Golub, M. (1999), „Adaptive genetic algorithm", ´ Journal of computing and information technology 7(3), 229–235.
23. Cover, T. M. and Thomas, J. A. (1991), „Entropy, relative entropy and mutual information", Elements of information theory 2, 1–55.
24. Paul, T. K. and Iba, H. (2005), Extraction of informative genes from microarray data, in „Proceedings of the 7th annual conference on Genetic and evolutionary computation", pp. 453–460.
25. Li, J., Tang, X., Zhao, W. and Huang, J. (2007), „A new framework for identifying differentially expressed genes", Pattern Recognition 40(11), 3249–3262.
26. Battiti, R. (1994), „Using mutual information for selecting features in supervised neural net learning", IEEE Transactions on neural networks 5(4), 537–550.
27. Zhu, Z., Ong, Y.-S. and Dash, M. (2007), „Markov blanket-embedded genetic algorithm for gene selection", Pattern Recognition 40(11), 3236–3248.
28. Dettling, M. and Buhlmann, P. (2002), „Supervised clustering of genes", ¨ Genome biology 3(12), research0069–1.
29. D´ıaz-Uriarte, R. and De Andres, S. A. (2006), „Gene selection and classification of microarray data using random forest", BMC bioinformatics 7(1), 3.
30. Cilimkovic, M. (2015), „Neural networks and back propagation algorithm", Institute of Technology Blanchardstown, Blanchardstown Road North Dublin 15.
31. Rong, H.-J., Ong, Y.-S., Tan, A.-H. and Zhu, Z. (2008), „A fast pruned-extreme learning machine for classification problem", Neurocomputing 72(1-3), 359–366.
32. Gu, B., Sheng, V. S., Tay, K. Y., Romano, W. and Li, S. (2014), „Incremental support vector learning for ordinal regression", IEEE Transactions on Neural networks and learning systems 26(7), 1403–1416.
33. Gu, B. and Sheng, V. S. (2016), „A robust regularization path algorithm for ν-support vector classification", IEEE Transactions on neural networks and learning systems 28(5), 1241–1248.
34. Reunanen, J. (2003), „Overfitting in making comparisons between variable selection methods", Journal of Machine Learning Research 3(Mar), 1371–1382.
35. Somol, P., Pudil, P., Novovicov ˇ a, J. and Paclık, P. (1999), „Adaptive floating search ´ methods in feature selection", Pattern recognition letters 20(11-13), 1157–1163