MF-GARF: Hybridizing Multiple Filters and GA Wrapper for Feature Selection of Microarray Cancer Datasets

Pakizah Saqib*, Usman Qamar*, Reda Ayesha Khan*, Andleeb Aslam*

*Department of Computer and Software Engineering, CEME,
National University of Sciences and Technology (NUST), Islamabad, Pakistan

pakizahfatima@gmail.com, usmanq@ceme.nust.edu.pk, reda.ayesha@gmail.com, andleebaslam0@gmail.com

***Abstract*— DNA Microarray technology is a valuable advancement in medical field but it gives birth to many challenges like curse of dimensionality, storage and computational requirements. In this paper we have proposed, a multiple filters and GA wrapper based hybrid approach (MF-GARF) that incorporates Random forest as fitness evaluator of features. The proposed hybrid approach MF-GARF is comprised of three phases relevancy block; containing information theory based filters Information Gain, Gain Ratio and Gini Index, responsible for ensuring relevancy and removal of irrelevant and noisy features. Second phase is Redundancy block; incorporating Pearson Correlation statistics to remove redundancy among features, and then final phase Optimization Block; containing Genetic Algorithm wrapper with Random Forest as fitness evaluator, responsible for generating an optimal feature subset with high predictive power. Random Forest with 10-fold cross validation is used to calculate the classification accuracy of selected feature subset. Experiments are carried out on 7 publically available benchmark Microarray cancer datasets and the proposed algorithm has achieved good accuracy with minimal selected features for all datasets. The comparison with other state of the art hybrid techniques validates the effectiveness of our proposed approach.**

***Keywords*— Feature Selection, Gene Selection, Hybrid, Genetic Algorithm, Random Forests, Filters, Microarray Cancer Datasets**

I. Introduction

Data is too diverse. Diversity of data has made feature selection a fundamental step for many data mining tasks, especially for processing of high dimensional data like microarray datasets comprising of more than thousands of features and small set of samples. The rapid growth of data gives birth to many challenges like curse of dimensionality, storage and computational requirements. Gathering data is not a problem but obtaining meaningful information from raw data is critically important. The abundance of data demands optimal and efficient algorithms to process raw data to retrieve useful information[1].

DNA Microarray technology is a valuable advancement in medical field that facilitates medical specialists in monitoring and profiling gene expressions of an organisms. With the help of this technology, biologist can profile thousands of gene expressions in a single experiment[2]. DNA chip can profile thousands of genes, but not all gene expressions contribute to the diagnosis of a disease. Microarray gene expression dataset contains plenty of irrelevant and redundant genes that may halt the process of correct diagnosis of a disease.

The process of analyzing the gene expression dataset to find out the most informative genes from the pool of noisy, redundant and irrelevant is known as Gene expression analysis. Analysis of microarray gene expression datasets is a crucial task to resolve the issues and challenges associated with them. Classification of microarray gene expressions is a NP hard problem, it contains thousands of genes and small sample size, that gives birth to the issue of curse of dimensionality. To deal with the issue of high dimensionality and low sparsity, dimensionality reduction is one solution. Dimensionality reduction in terms of feature (gene) selection is of great interest as large number of features (genes) and small sample size leads to overfitting of model, poor model learning, erroneous predictions. Moreover, model construction and learning over such dataset are computationally expensive and inefficient.

Feature selection [2][3] is an effective way to solve the curse of dimensionality of microarray datasets. It is a common pre-processing technique in data mining tasks to enhance the efficiency of classifiers. Generally, Feature selection strategies are classified into three categories, filters, wrapper and hybrid. Filters select features by evaluating each feature individually and scoring statistically without using the heuristics of any classifier. Wrapper evaluate features using performance accuracy of the classifiers and are more efficient in terms of performance than filter but are computationally expensive. Hybrid is a combination of any of the two or more feature selection approaches to overcome the issues associated with the individual approach and merge the goodness of the combined approaches.

II. Literature Work

In this section we have discussed the existing state of the art filter, wrapper and hybrid approaches for supervised feature selection of microarray caner datasets.

A. *Filters approaches for feature selection*

Filter techniques because of it generalization properties have been used by many researchers for feature selection of microarray cancer datasets. The recent researches has presented many novel filter approaches to rank features like Hidden Markov's Model (HMM) [4], X- variance[5] Mutual congestion[5], Qualitative Mutual Information[6] used Mutual information (MI) with Random forest feature importance, Pareto based feature Ranking technique [8] for multi-objective optimization, Partial Maximum Correlation Information (PMCI) [9], a multiple synergy filter based feature selection approach that assesses feature importance by extracting orthogonal components from feature space. Correlation based feature selection[10], mRMR[11] approach that covers both the relevancy and redundancy in parallel manner. Filter based Feature selection has more generalization properties as compared to other approaches but they lack the capabilities to reduce the dimensions in case of high dimensional datasets and thus do not generate the good prediction accuracies. To overcome the drawbacks associated with filter approaches, wrapper and hybrid approaches are proposed that involve the heuristics of classifiers to evaluate the performance of selected features.

B. Wrapper Approaches for Feature Selection

In wrapper approach, search process is wrapped around an induction algorithm usually a classifier. And, the performance accuracy or classification error rate of the classifier is used to evaluate the selection of best feature subset. Wrapper are more efficient than filters in terms of performance as it considers the correlation among features and directly incorporates the biases of induction algorithm. Many studies has proposed wrapper based approaches for feature selection of microarray cancer datasets like PSO-SVM [12], GA- SVM [12], ABC-SVM[12], FF-SVM[13], ACO-SVM[14], HS-GA[15], BPSO-CGA[16], and HPSO-LS[17]. Wrapper methods are better alternative to filters for supervised learning problems being efficient in performance but are computationally expensive, hence require plenty of computational resources for high dimensional datasets. Moreover, wrapper models are prone to overfitting, calling classifier again and again for the evaluation of each feature subset results in overfitting.

C. Hybrid Approaches for Feature Selection

Hybrid approach for feature selection is either combination of same or two or more different techniques, with an aim to combine the best traits of combined feature selection approaches. The most common hybrid combination is of filter and wrapper. In which filters are generally employed as a pre-processor for the later stage i.e. wrapper, as it statistically scores the features and pool out the informative genes using less computational resources. Features with high ranking or scores are saved and used as initial search space for wrappers, that uses the heuristic of learning algorithms to calculate the prediction accuracies of different feature candidate subsets and opt out the most accurate candidate subset. Practically, any combination of filter and wrapper can be used for constructing a hybrid but in literature variety of novel and efficient combinations of filter and wrapper has been suggested. Most of researchers have employed bio- inspired evolutionary method as wrappers like

Table 1. Summary of Existing Hybrid Feature Selection Approaches for Microarray Datasets

Proposed Hybrid Approach	**Classifier**	**Performance Validation**
Fisher Score and AI tuned Genetic Algorithm (IDGA – F) [18]	▪ SVM ▪ kNN ▪ NB	LOOCV
Laplacian Score and AI tuned Genetic Algorithm (IDGA – L) [18]	▪ SVM ▪ kNN ▪ NB	LOOCV
Information Gain and Genetic Algorithm (IG-GA)[19]	▪ Genetic Programming (GP)	10 Fold Cross Validation
Mutual Information (MI) and Adaptive Stem Cell Optimization (ASCO)[21]	▪ Fuzzy Classifier	Classification Accuracy of Rules Generated
Fisher Criterion and Cellular Learning Algorithm with Ant Colony Optimization (CLACOFS)[20]	▪ NB ▪ kNN ▪ SVM	ROC curve
Independent Component Analysis and Artificial Bee Colony Optimization (ICA + ABC)[23]	▪ Naïve Bayesian (NB)	LOOCV
Fisher Score Criterion and Multi-objective Binary Bat Algorithm with Local Search (FS-MOBBA-LS)[22]	▪ SVM ▪ NB ▪ kNN	10 Fold Cross Validation
Minimum Redundancy and Maximum Relevance with Artificial Bee Colony Optimization (mRMR - ABC)[24]	▪ SVM	LOOCV
Mutual Information Maximization and Adaptive Genetic Algorithm (MIMAGA)[25]	▪ ELM ▪ RELM ▪ BP ▪ SVM	Classification Accuracy
Random Forest Ranking and Binary Black Hole Algorithm (RFR-BBHA)[26]	▪ Bagging	10 Fold Cross Validation
Fisher Markov Selector and Multi-objective binary biogeographic based optimization (MOBBBO)[27]	▪ SVM	LOOCV
Symmetrical Uncertainty and Harmonic Search Algorithm (SU-HSA) [1]	▪ Instance Based Classifier ▪ NB	10 Fold Cross Validation
Fast Correlation Based Filter and Genetic Algorithm (FCBF-GA) [29]	▪ SVM	LOOCV
Fast Correlation Based Filter and Particle Swarm Optimization (FCBF-PSO) [29]	▪ SVM	LOOCV

Genetic Algorithm [18],[19],[25],[30] Ant Colony Optimization [23], Bat Algorithm [22] and Artificial Bee Colony Optimization [24]. The Table 1 summarises the proposed hybrid approaches in recent years for feature selection of microarray datasets along with classifiers and performance validation metrics.

III. Proposed Framework

In this section, we present a hybrid approach for feature selection of microarray cancer dataset that preserves the advantages of filter and wrapper methods while mitigating their drawbacks. A schema of proposed framework is given in Figure 1.

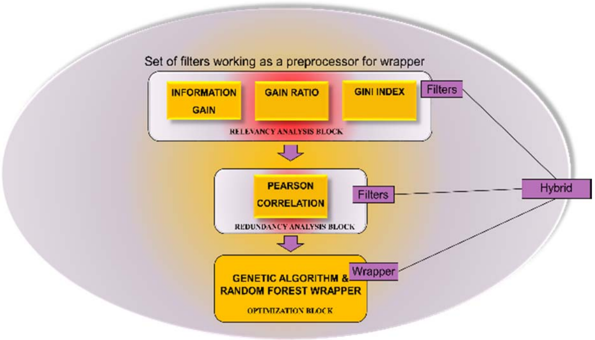

Figure 1 – Schema of Proposed Framework MF-GARF

In the first stage, we have used a set of three information based filter techniques Information Gain, Gain Ratio and Gini Index, each of these filter technique score each feature statistically without any learning algorithm and selects the top-scoring features filtered by each filter method, meeting a specific threshold criterion. A feature set is then created by taking union of features opted by each filtering technique. All three filters rank feature based on information they add to the class label, so directly ensure the relevancy of selected feature to the class label.

Another filtering technique Pearson Correlation is used that removes the redundancy from the selected features. Thus turning a high dimensional dataset into a small amount of feature pool, serving as a reduced search space for an optimal wrapper approach Genetic Algorithm that incorporates the Random Forest to evaluate the fitness of each selected feature subset. We have used set of univariate filters that score each feature individually thus do not consider the relationship among feature, the subset of feature may bring more information to the leaning model instead of an isolated feature but this may induce redundancy.

Feature subset selected by filters can be still large and it's not tuned to any classifier that's why we introduced a second stage where a wrapper is used to reduce the dimensionality of the feature subset. Motivation of this hybrid approach is to involve both important aspects of feature selection i.e. relevancy and redundancy analysis of features. And bring forth an optimal subset of features. Wrapper Approach are computationally expensive for high dimensional datasets, that why in the initial stage we have used filters that serve as a pre-processing step for wrapper that reduces its search space. Thus this approach of feature selection along with high dimensionality reduction provide a time and space complexity improvement.

A. Relevancy Analysis Block

In this block, relevancy of each feature (gene) with the target class is calculated using information FEF based univariate filters Information gain, Gini index, and Gain Ratio.

1) Information Gain: Information gain is one of the most preferred feature ranking filter that measure the relevancy of each feature and helps to make decision either the feature should be chosen or not. Information gain is a symmetrical measure of mutual dependence between two variables. It captures information regarding one random variable, through other random variable. It is one of the variants used by decision tree in machine learning to capture the importance of features. Information gain is based on entropy, which is measure of randomness in the information being processed, where there is a minimum entropy there is a maximum information gain. For each feature information gain value is calculated. Greater value of Information gain depicts relevancy of feature to the target class. A threshold criterion is adjusted to make a choice of features to be kept, a feature with information gain value above or equal to threshold value are kept while others are discarded [30]. The Entropy (E) and Information Gain (IG) is calculated using Eq. (1) an (2) respectively.

$$E = -\sum_{i=1}^{c}(p_i \, log_2(p_i)) \tag{1}$$

$$IG\left(D_{P,}f\right) = E\left(D_{P,}\right) - \frac{N_{left}}{N}\,E\left(D_{left}\right) - \frac{N_{right}}{N}\,E\left(D_{right}\right) \tag{2}$$

2) Gain Ratio: Gain Ratio [31], a term coined by Ross Quinlan, is an improved version of Information Gain. It scores the features in a similar way as the Information gain, calculates the information regarding one random variable (x), through other random variable. But Gain ratio involves intrinsic information in order to give overall score to each feature. Information gain favours multi-valued features. The approach of gain ratio is to amplify the information gain while limiting the number of its values. The equation for gain ratio is given in Eq. (3) follow: -

$$Gain\ Ratio(x) = \frac{\text{Information Gain (x)}}{\text{Intrinsic Value (x)}} \tag{3}$$

3) Gini Index: Gini Index [32] is a term coined by Corrado Gini, an Italian statistician in 1912. It measures the impurity and uncertainty among the values of the features. Greater the value of an index, more the data will be distributed. This measure of impurity is used by Classification and Regression Tree (CART). This technique is considered to be more fast as compared to other filtering technique. We have used Gini index to evaluate the goodness of each feature. The basic purpose behind using this approach is to select those features having minimum Gini index value and thus bringing in maximum purity improvement. The feature with minimum Gini index value is considered as a splitting point in CART as it brings maximum purity. The equation for the Gini Index is as follow: -

$$Gini = 1 - \sum_{i=1}^{c}(p_i)^2 \qquad (4)$$

For each filter approach following steps are followed:

1. Microarray cancer dataset is retrieved.
2. Missing values are removed.
3. Dataset is passed to filter that evaluates the feature according to its evaluation criteria, and evaluation score is assigned to each feature.
4. Evaluation Scores are normalized using formula.
5. Now feature scores are compared with threshold criterion i.e. *th >= 0.5*. Feature satisfying the threshold criteria are kept while others are discarded.

At final step, all the feature sets are merged into a unified set that serves as an input for redundancy analysis block. The Figure 2 and 3 give the flowchart for proposed algorithm.

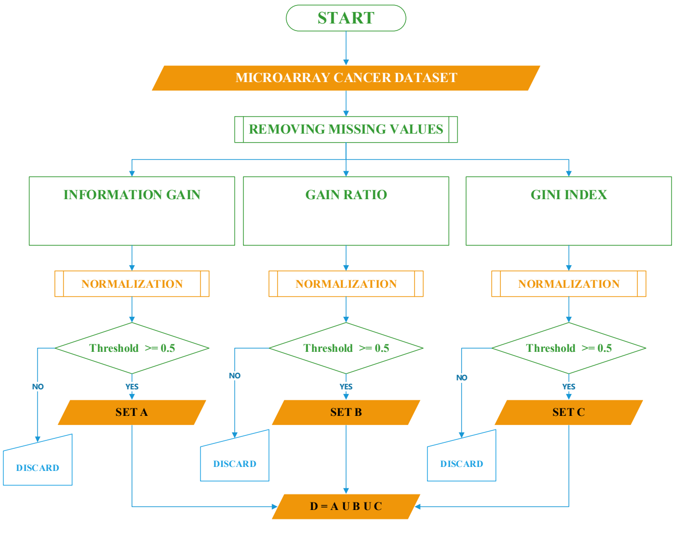

Figure 2 - Flowchart for Relevancy Analysis Block

B. Redundancy Analysis Block

The above used filter approaches are univariate, they evaluate and score features independently without considering feature interaction, so there are chances of presence of highly correlated features that induce redundancy in data and add no additional information to the model, hence these features are undesirable. For that we have incorporated Pearson Correlation coefficient measure to calculate the correlation among features.

1) Pearson Correlation: Pearson Correlation [33] has been used to overcome the issue of redundancy among the features. The aim is to come up with a feature subset incorporating features that are highly correlated with the target class but have low correlation with each other. So using Pearson correlation, that is one of the most helpful statistically measure for figuring out the strength and relationship among variable, correlated features are removed as they do not add any additional information to the learning model. Individually they might have some presence but there exist other features similar in behaviour, having same impact on prediction, thus resulting in redundancy. Removing correlated features saves space and time of calculation of complex algorithms. Moreover, it also makes processes easier to design, analyse, understand and comprehend. The equation for Pearson Correlation (r) is as follow: -

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}} \qquad (5)$$

Features with high correlation value i.e. greater and equal to 0.85 with other features are removed from the set D, and new set D’ is obtained.

The set D’ serves as search space of Genetic algorithm wrapper in optimization block.

C. Optimization Block

The optimization block is comprised of genetic algorithm wrapper with random forest as an induction algorithm to bring forth an optimal set of features (genes).

1) Genetic Algorithm: Genetic Algorithm based on Charles Darwin theory of natural evolution, follows the pattern of natural selection to select the individuals with maximum fitness score to reproduce offspring for the next generation. Here from fitness score of individual we mean those individual that have more power to withstand the environmental changes. Genetic algorithm (GA) [18] being one of the most promising and efficient optimization technique has been used in combination with many wrapper and hybrid methods for feature selection and classification of high dimensional dataset, especially microarray datasets, for last two decades.

Genetic Algorithm [18], [29] performs a refined feature selection from a pool of highly informative features. Genetic Algorithm is a global search technique that improve the quality of selected feature by finding an optimal feature subset. It looks into the search space for the fittest feature subset that produces the best classification accuracy. The initial population, fitness function, selection, crossover and mutation operator are the five main components of the genetic algorithm. The population of Genetic Algorithm is comprised of chromosomes, and each chromosome in population corresponds to a solution to the optimization problem. Each chromosome is incorporated as a binary sequence (0’s and 1’s). The length of chromosome corresponds to the number of features in dataset, the presence of feature is represented by 1, while 0 indicates the absence of feature. And role of each chromosome for the next generation is determined by the fitness value it acquires. In our proposed hybrid approach, the fitness is measured as a function of the accuracy of the Random Forest classifier with which GA is wrapped. For fitness evaluation of each feature subset Random Forest, an ensemble model, is used. It has been discussed in later section.

2) Random Forest Classifier: Random Forest [7],[26],[35] operates as an ensemble approach it uses decision tree as a base model. The best thing about Random forest is that it considers “the wisdom of crowd”. It builds multiple uncorrelated decision trees, each decision tree individually predicts the class, final prediction of class is based on the majority vote. In our proposed approach we have used random forest classifier as an induction algorithm for GA wrapper. Here random forest has used Gini index as a splitting criteria, as it does not involve logarithm thus making random forest computationally inexpensive.

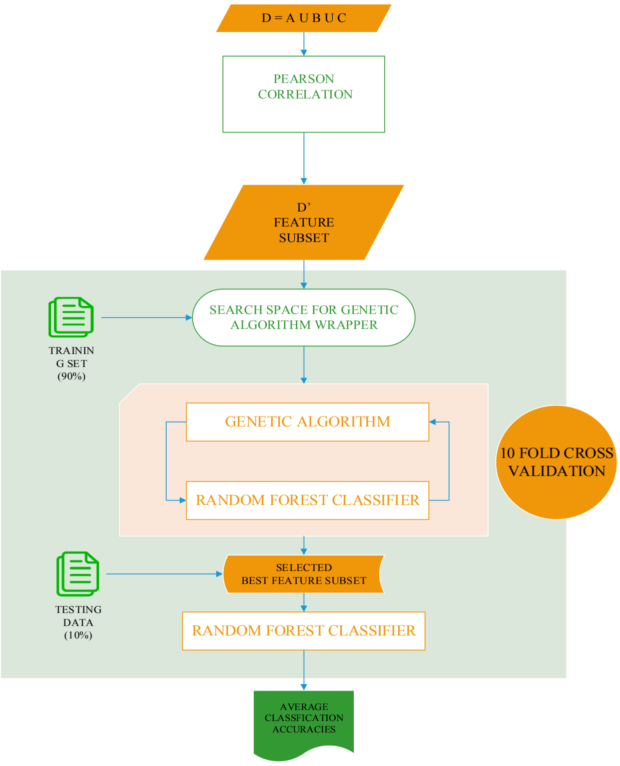

Figure 3 - Flowchart of Redundancy Analysis and Optimization Block

IV. Experimental Setup and Results

For evaluation of proposed frameworks, the algorithm is implemented in python. And all the experiments are carried out on HP notebook with 2.7 GHz processor speed and 4GB RAM. Moreover, datasets used for evaluation of proposed framework, the parameter tuning of wrapper, classifier, and experimental results are discussed in this section.

A. Datasets

The algorithm is evaluated on 7 publically available binary class microarray cancer datasets [34][36] including Colon , Leukemia, DLBCL, Ovarian, CNS, Prostate and Breast Cancer datasets. The description of dataset is given in Table 2.

Table 2. Description of Microarray Cancer Datasets

Datasets	Attributes	Instances	Classes
Colon	2000	62	Normal / Tumor
Leukemia	7129	72	ALL / AML
DLBCL	4026	55	DLBCL/FL
Ovarian	15155	253	Normal/Cancer
CNS	7129	60	1/0
Prostate	12533	102	Normal/Tumor
Breast	24481	97	Relapse/Non-Relapse

B. Parameter Tunning

For experimentation, few parameters are tuned. In relevancy analysis block, the threshold for maintaining relevancy is set to 0.5 for all three filters. In redundancy analysis block, Correlation value is set to 0.85 to extract the feature having same or above correlation value with other features. The parameters of Genetic algorithm i.e. population is set to 80 for Prostate and Breast cancer datasets, while for rest of the datasets, its value is set to 20. Number of generations is set to 30 to iterate the process 30 times to get the most optimal subset of features. The values of probability of crossover (Pc) is set to 0.5 while probability of mutation (Pm) is assigned 0.01 value. The parameters of random forest classifier number of trees, maximum depth and splitting criterion are set to 25, 25 and Gini Index respectively as suggested by study [35]. The experiment is repeated for 10 times using 10-fold cross validation to get the accurate result mitigating any chances of model overfitting. Random forest classifier is used to calculate the accuracy of feature subset. Moreover, precision, recall and AUC are also computed.

C. Experimental Results

In this section, the experimental results are discussed, the Table 3 gives the feature count, while Table 4 gives classification accuracies obtained at the end of each phase of our proposed algorithm to give an idea how it works.

Table 3. Feature Count After Each Stage of Proposed Algorithm

Dataset	**Raw Feature Count**	**HYBRID**		
		Relevancy Analysis Block	**Redundancy Analysis Block**	**Optimization Block**
Colon	2000	58	34	3
Leukemia	7129	112	108	3
DLBCL	4026	384	334	4
Ovarian	15155	59	20	5
CNS	7129	1074	257	7
Prostate	12533	90	45	5
Breast	24881	8511	2593	10

Table 4. Classification Accuracies after each stage of proposed algorithm

Dataset	**Raw Accuracy**	**HYBRID**		
		Relevancy Analysis Block	**Redundancy Analysis Block**	**Optimization Block**
Colon	63.16%	89.47%	89.47%	96.77%
Leukemia	90.48%	95.24%	100%	100%
DLBCL	98.1%	99.05%	100%	100%
Ovarian	96.05%	98.68%	98.68%	100%
CNS	66.67%	66.67%	72.22%	93.33%
Prostate	70.97%	93.55%	96.77%	98.04%
Breast	58.62%	72.41%	86.21%	86.60%

The Table 5 shows the accuracy, precision, recall and AUC for all datasets. As the result shows, the proposed algorithm has given 100% accuracies for Leukemia, DLBCL and Ovarian datasets.

Table 5. Accuracy, Precision, Recall, AUC

Dataset	Avg. Accuracy	Avg. Precision	Avg. Recall	Avg. AUC
Colon	96.77%	100%	90.91%	95.1%
Leukemia	100%	100%	100%	100%
DLBCL	100%	100%	100%	100%
Ovarian	100%	100%	100%	100%
CNS	93.3%	92.68%	97.43%	90.6%
Prostate	98.04%	98.08%	98.08%	97.3%
Breast	86.60%	88.54%	88.24%	85.4%

V. Comparison with State of the Art Techniques

This section shows the comparison of our proposed hybrid approach with other state of the art hybrid approaches in terms of classification accuracies and number of selected genes to validate the effectiveness of our proposed algorithm (MF-GARF). The comparison is presented in Table 6. In most of the hybrid techniques bio-inspired evolutionary wrappers are used in combination with filters so it would be a fair comparison. The results show that proposed algorithm (MF-GARF) has outperformed other hybrid approaches in 5 out of 7 microarray cancer datasets. The proposed algorithm has achieved 100% accuracies for 3 datasets Leukemia, DLBCL and Ovarian cancer dataset with only 3,4, and 5 selected features. In IG-GA [19], only information gain is used as a pre-processer for GA wrapper while we have employed multiple filters along with GA wrapper to reduce the space complexity and have achieved more accurate classification accuracies for Colon, Leukemia, DLBCL, and CNS comparatively with lesser no. of selected features. For prostate dataset CLACOFS [20] and ICA + ABC [23] has achieved better accuracies than our proposed approach. For Leukemia, MF-GARF has overshadowed the performance of other state of the art techniques. For CNS and Breast Cancer dataset, MF-GARF has achieved 93.33% and 86.60% accuracy which is higher than classification accuracies of other techniques.

VI. Conclusions

To overcome the high dimensionality and low sparsity issue of feature selection, we have proposed a hybrid feature selection approach (MF-GARF) based on multiple filters and Genetic Algorithm (GA) wrapper in combination with Random Forest (RF) classifier. The experimentation and results show that proposed algorithm achieves higher accuracies than other state of the art techniques presented in literature. In this study we have employed the information based filters that performs evaluation of features based on the amount of information they provide about the target class. In future work, we can use filters employing variable feature evaluation criteria like distance, similarity, or consistency. Moreover, ensemble approach instead of hybridization can be incorporated at pre-processing stage to retain only those features voted by majority of feature evaluation techniques.

References

[1] J. Li and H. Liu, "Challenges of Feature Selection for Big Data Analytics", IEEE Intelligent Systems, vol. 32, no. 2, pp. 9-15, 2017. doi: 10.1109/mis.2017.38.

[2] D. Rew, "DNA microarray technology in cancer research", European Journal of Surgical Oncology (EJSO), vol. 27, no. 5, pp. 504-508, 2001. doi: 10.1053/ejso.2001.1116.

[3] G. Chandrashekar and F. Sahin, "A survey on feature selection methods", Computers & Electrical Engineering, vol. 40, no. 1, pp. 16-28, 2014. doi: 10.1016/j.compeleceng.2013.11.024.

[4] J. Ang, A. Mirzal, H. Haron and H. Hamed, "Supervised, Unsupervised, and Semi-Supervised Feature Selection: A Review on Gene Selection", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 13, no. 5, pp. 971-989, 2016. doi: 10.1109/tcbb.2015.2478454.

[5] M. Momenzadeh, M. Sehhati and H. Rabbani, "A novel feature selection method for microarray data classification based on hidden Markov model", Journal of Biomedical Informatics, vol. 95, p. 103213, 2019. doi: 10.1016/j.jbi.2019.103213.

[6] M. Alirezanejad, R. Enayatifar, H. Motameni and H. Nematzadeh, "Heuristic filter feature selection methods for medical datasets", Genomics, 2019. doi: 10.1016/j.ygeno.2019.07.002.

[7] A. Nagpal and V. Singh, "A Feature Selection Algorithm Based on Qualitative Mutual Information for Cancer Microarray Data", Procedia Computer Science, vol. 132, pp. 244-252, 2018. doi: 10.1016/j.procs.2018.05.195.

[8] R. Dash, "A two stage grading approach for feature selection and classification of microarray data using Pareto based feature ranking techniques: A case study", Journal of King Saud University - Computer and Information Sciences, 2017. doi: 10.1016/j.jksuci.2017.08.005.

[9] M. Yuan, Z. Yang and G. Ji, "Partial maximum correlation information: A new feature selection method for microarray data classification", Neurocomputing, vol. 323, pp. 231-243, 2019. doi: 10.1016/j.neucom.2018.09.084.

[10] A. Hasan and Md. Akhtaruzzaman Adnan, "High dimensional microarray data classification using correlation based feature selection," 2012 International Conference on Biomedical Engineering (ICoBE), Penang, 2012, pp. 319-321. doi: 10.1109/ICoBE.2012.6179029.

[11] M. Mandal and A. Mukhopadhyay, "An Improved Minimum Redundancy Maximum Relevance Approach for Feature Selection in Gene Expression Data", Procedia Technology, vol. 10, pp. 20-27, 2013. doi: 10.1016/j.protcy.2013.12.332.

[12] H. Alshamlan, G. Badr and Y. Alohali, "ABC-SVM: Artificial Bee Colony and SVM Method for Microarray Gene Selection and Multi Class Cancer Classification", International Journal of Machine Learning and Computing, vol. 6, no. 3, pp. 184-190, 2016. doi: 10.18178/ijmlc.2016.6.3.596.

[13] N. Almugren and H. Alshamlan, "FF-SVM: New FireFly-based Gene Selection Algorithm for Microarray Cancer Classification," 2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Siena, Italy, 2019, pp. 1-6. doi: 10.1109/CIBCB.2019.8791236

[14] J. C. H. Hernandez, B. Duval and J.-K. and Hao, "A Genetic Embedded Approach for Gene Selection and Classification of Microarray Data," in Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, 2007. doi: 10.1007/978-3-540-71783-6_9

[15] S. Vijay and P. GaneshKumar, "Fuzzy Expert System based on a Novel Hybrid Stem Cell (HSC) Algorithm for Classification of Micro Array Data", Journal of Medical Systems, vol. 42, no. 4, 2018. doi: 10.1007/s10916-018-0910-0.

[16] L. Chuang, C. Yang, J. Li and C. Yang, "A Hybrid BPSO-CGA Approach for Gene Selection and Classification of Microarray Data", Journal of Computational Biology, vol. 19, no. 1, pp. 68-82, 2012. doi: 10.1089/cmb.2010.0064.

[17] P. Moradi and M. Gholampour, "A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy", Applied Soft Computing, vol. 43, pp. 117-130, 2016. doi: 10.1016/j.asoc.2016.01.044.

Table 6. Comparison with State of the Art Hybrid Techniques (Number of selected features is written in parenthesis)

Feature Selection Algorithm	Classifier	Colon	Leukemia	DLBCL	Ovarian	CNS	Prostate	Breast
(IDGA – F) [18]	SVM	-	98.1(12)	98.1(10)	-	-	96.8(25)	-
	kNN	-	98.1(14)	95.4(8)	-	-	92.5(29)	-
	NB	-	95.2(18)	97.9(9)	-	-	92.3(30)	-
(IDGA – L) [18]	SVM	-	95.6(13)	99.7(21)	-	-	89.3(32)	-
	kNN	-	97.4(8)	91.2(23)	-	-	73.6(33)	-
	NB	-	97.7(8)	93.0(20)	-	-	56.7(25)	-
(IG-GA) [19]	GP	85.48(60)	97.06(3)	94.87(110)	-	86.67(38)	100(30)	-
(MI - ASCO) [21]	Fuzzy Classifier	**98.25(NAN)**	94.75(NAN)	-	-	-	84.68(NAN)	-
(CLACOFS) [20]	NB	-	97.60(6)	-	-	-	99.10(7)	-
	kNN	-	95.95(4)	-	-	-	99.85(15)	-
	SVM	-	95.95(3)	-	-	-	98.35(14)	-
(ICA + ABC) [23]	NB	**98.14(16)**	98.68(12)	-	-	-	98.88(16)	-
(FS-MOBBA-LS) [22]	SVM	-	97.1(3)	-	-	-	94.1(6)	-
	NB	-	**100(3)**	-	-	-	97.1(6)	-
	kNN	-	**100(3)**	-	-	-	97.1(6)	-
(mRMR - ABC)[24]	SVM	**96.77(15)**	100(14)	-	-	-	-	-
(MIMAGA) [25]	ELM	89.09(7)	97.62(19)	-	-	-	96.54(3)	82.47(6)
(RFR-BBHA)[26]	Bagging	91.93(4)	-	-	-	86.66(2)	-	-
(SU-HSA)[1]	IB1	87.15(22)	99.53(23)	-	99.94(15)	-	-	83.39(24)
	NB	87.53(9)	100(26)	-	99.65(12)	-	-	75.97(15)
(FCBF-GA)[29]	SVM	96.30(1536)	-	100(2330)	-	-	-	-
(FCBF-PSO)[29]	SVM	96.30(999)	-	100(3204)	-	-	-	-
Proposed MF-GARF	RF	96.77(4)	**100(3)**	**100(4)**	**100(5)**	**93.33(7)**	**98.04(5)**	**86.60(10)**

[18] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts", Genomics, vol. 109, no. 2, pp. 91-107, 2017. doi: 10.1016/j.ygeno.2017.01.004.

[19] H. Salem, G. Attiya and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles", Applied Soft Computing, vol. 50, pp. 124-134, 2017. doi: 10.1016/j.asoc.2016.11.026.

[20] F. Vafaee Sharbaf, S. Mosafer and M. Moattar, "A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization", Genomics, vol. 107, no. 6, pp. 231-238, 2016. doi: 10.1016/j.ygeno.2016.05.001.

[21] S. Vijay and P. GaneshKumar, "Fuzzy Expert System based on a Novel Hybrid Stem Cell (HSC) Algorithm for Classification of Micro Array Data", Journal of Medical Systems, vol. 42, no. 4, 2018. doi: 10.1007/s10916-018-0910-0.

[22] M. Dashtban, M. Balafar and P. Suravajhala, "Gene selection for tumor classification using a novel bio-inspired multi-objective approach", Genomics, vol. 110, no. 1, pp. 10-17, 2018. doi: 10.1016/j.ygeno.2017.07.010.

[23] R. Aziz, C. Verma and N. Srivastava, "A novel approach for dimension reduction of microarray", Computational Biology and Chemistry, vol. 71, pp. 161-169, 2017. doi: 10.1016/j.compbiolchem.2017.10.009.

[24] H. Alshamlan, G. Badr and Y. Alohali, "mRMR-ABC: A Hybrid Gene Selection Algorithm for Cancer Classification Using Microarray Gene Expression Profiling", BioMed Research International, vol. 2015, pp. 1-15, 2015. doi: 10.1155/2015/604910.

[25] I. Jain, V. Jain and R. Jain, "Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification", Applied Soft Computing, vol. 62, pp. 203-215, 2018. doi: 10.1016/j.asoc.2017.09.038.

[26] E. Pashaei, M. Ozen and N. Aydin, "Gene selection and classification approach for microarray data based on Random Forest Ranking and BBHA," 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Las Vegas, NV, 2016, pp. 308-311. doi: 10.1109/BHI.2016.7455896

[27] X. Li and M. Yin, "Multiobjective Binary Biogeography Based Optimization for Feature Selection Using Gene Expression Data", IEEE Transactions on NanoBioscience, vol. 12, no. 4, pp. 343-353, 2013. doi: 10.1109/tnb.2013.2294716.

[28] S. Shreem, S. Abdullah and M. Nazri, "Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm", International Journal of Systems Science, vol. 47, no. 6, pp. 1312-1329, 2014. doi: 10.1080/00207721.2014.924600.

[29] H. Djellali, S. Guessoum, N. Ghoualmi-Zine and S. Layachi, "Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection," 2017 5th International Conference on Electrical Engineering - Boumerdes (ICEE-B), Boumerdes, 2017, pp. 1-6. doi: 10.1109/ICEE-B.2017.8192090

[30] F. He, H. Yang, Y. Miao and R. Louis, "A hybrid feature selection method based on genetic algorithm and information gain," 2016 5th International Conference on Computer Science and Network Technology (ICCSNT), Changchun, 2016, pp. 320-323. doi: 10.1109/ICCSNT.2016.8070172

[31] N. Prasad and M. M. Naidu, "Gain Ratio as Attribute Selection Measure in Elegant Decision Tree to Predict Precipitation," 2013 8th EUROSIM Congress on Modelling and Simulation, Cardiff, 2013, pp. 141-150. doi: 10.1109/EUROSIM.2013.35

[32] S. Sivagama Sundhari, "A knowledge discovery using decision tree by Gini coefficient," 2011 International Conference on Business, Engineering and Industrial Applications, Kuala Lumpur, 2011, pp. 232-235. doi: 10.1109/ICBEIA.2011.5994250

[33] E. Reggiani, E. D'Arnese, A. Purgato and M. D. Santambrogio, "Pearson Correlation Coefficient Acceleration for Modeling and Mapping of Neural Interconnections," 2017 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Lake Buena Vista, FL, 2017, pp. 223-228. doi: 10.1109/IPDPSW.2017.63

[34] Bioinformatics Laboratory", Biolab.si, 2019. [Online]. Available: http://www.biolab.si/supp/bi-cancer/projections/info/prostata.html. [Accessed: 04- Aug- 2019].

[35] P. Saqib, U. Qamar, A. Aslam and A. Ahmad, "Hybrid of Filters and Genetic Algorithm-Random Forests Based Wrapper Approach for Feature Selection and Prediction," in *Intelligent Computing - Proceedings of the Computing Conference, CompCom 2019*, London, UK, 2019. doi: 10.1007/978-3-030-22868-2_15

[36] 2019. [Online]. Available: http://csse.szu.edu.cn/staff/zhuzx/Datasets.html. [Accessed: 01- Aug-2019]

Pakizah Saqib received the B.S. degree in Software Engineering from PUCIT, University of the Punjab, Lahore, Pakistan. She is currently pursuing the M.S. degree in computer software engineering with the Computer and Software Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan. Her area of research is Gene expression dataset analysis using Machine Learning techniques.

Usman Qamar has over 15 years of experience in data engineering and decision sciences both in academia and industry having spent nearly 10 years in the UK. He has a Masters in Computer Systems Design from University of Manchester Institute of Science and Technology (UMIST), UK. His MPhil in Computer Systems was a joint degree between UMIST and University of Manchester which focused on feature selection in big data. In 2008/09 he was awarded PhD from University of Manchester, UK. His PhD specialization is in Data Engineering, Knowledge Discovery and Decision Science. His Post PhD work at University of Manchester, involved various research projects including hybrid mechanisms for statistical disclosure (feature selection merged with outlier analysis) for Office of National Statistics (ONS), London, UK, churn prediction for Vodafone UK and customer profile analysis for shopping with the University of Ghent, Belgium. He has also done a post-graduation in Medical & Health Research, from University of Oxford, UK, where he worked on evidence-based health care, thematic qualitative data analysis and healthcare innovation and technology.

He is director of Knowledge and Data Science Research Centre, a Centre of Excellence at NUST, Pakistan and principal investigator of Digital Pakistan Lab, which is part of National Centre for Big Data and Cloud Computing. He has authored over 150 peer reviewed publications which includes 2 books published by Springer & Co, 21 Book Chapters, 38 Impact Factor Journal Publications with a combined impact factor of 104 (Clarivate Analytics Impact Factor) and over 100 Conference Publications. Many of his papers have been awarded best research paper awards by Higher Education Commission, Pakistan. Because of his extensive publications he is member of Elsevier Advisory Panel. He has successfully supervised 4 PhD students and over 70 master students.

Dr. Usman has been able to acquire nearly PKR 100 million in research grants. He has received multiple research awards, including Best Book Award 2017/18 by Higher Education Commission (HEC), Pakistan, Best Researcher of Pakistan 2015/16 by Higher Education Commission (HEC), Pakistan, Best Overall NUST University Researcher Award 2016 and Best College of E&ME researcher award 2016 as well as gold in Research & Development category by Pakistan Software Houses Association (P@SHA) ICT Awards 2013 & 2017 and Silver award in APICTA (Asia Pacific ICT Alliance Awards) 2013 in category of R&D hosted by Hong Kong. He is also recipient of the prestigious Charles Wallace Fellowship 2016/17 as well as British Council Fellowship 2018, visiting research fellow at Centre of Decision Research, University of Leeds, UK and scientific director of Data and Text Mining Lab, Manchester Metropolitan. He is also an expert committee member of engineering & technology for the evaluation/recognition of national research journals for Higher Education Commission (HEC), Pakistan. Finally, he has the honour of being the finalist of the British Council's Professional Achievement Award 2016/17.

Reda A. Khan received the B.S. degree in Software Engineering from University of Engineering and Technology, Taxila, Pakistan. She is currently pursuing the M.S. degree in computer software engineering with the Computer and Software Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan. Her area of research is Ontology Engineering.

Andleeb Aslam received the B.S. degree in Software Engineering from University of Engineering and Technology, Taxila, Pakistan. She is currently pursuing the M.S. degree in computer software engineering with the Computer and Software Engineering Department, College of Electrical and Mechanical Engineering, National University of Sciences and Technology (NUST), Islamabad, Pakistan. Her area of research is Natural Language Processing (NLP).