Feature Selection Techniques for Cancer Classification applied to Microarray Data: A survey

Mohammed Qaraad
Department of Computer Science,Faculty of Science
Abdelmalek Essaadi University
Tetouan, Morocco
m.qaraad@gmail.com

Souad Amjad
Department of Computer Science,Faculty of Science
Abdelmalek Essaadi University
Tetouan, Morocco
amjd_souad@gmail.com

Hanaa Fathi
Department of Math and Computer Science,Faculty of Science
Menoufia University
Menoufia, Egypt
hanaa_4_ever@yahoo.com

Ibrahim I.M. Manhrawy
Department of Basic Sciences
Modern Academy for Engineering and Technology
Cairo, Egypt
ibrahimmanhrawy@gmail.com

***Abstract*— In multidimensional microarrays, data that collect gene expression profiles that fulfill the state of the cell at the molecular level. Feature selection and extraction have become an obvious need for the analysis of this microarray. There are many different methods for selecting and extracting attributes, and they are widely used. One of the serious tasks is to learn how to extract useful information from huge microarrays datasets with complex relationships between different genes. These methods are aimed at removing excess and irrelevant traits and extruding marker genes that effectively maintain classification accuracy. This report gives an overview of the various ways of performing dimensional reduction methods that were used in these microarrays to select important features and presents a comparison between them. The advantages and disadvantages of several methods are described in order to show an obvious idea of when to use each of them to save computational time and resources.**

Keywords*— *Microarray Data, Classification, Feature Selection, cancer, Machine Learning.

I. INTRODUCTION

During the last decade, the advent of DNA microarray technology has produced huge microarray data set, which motivated a new trace of research in both scientific fields machine learning and Bioinformatics[1]. This kind of data represents the gene expression profiles, that collected from tissue and cell samples, which could be beneficial for the gene-related studies that related to diagnosis disease or for determining distinct types of tumor [2]. Researchers have used these data to Analyze and classify different types of cancer or identify normal and patient people[3,4,5,6].

The gene expression data normally receive tens of thousands of genes for each data point (sample). Therefore, this data is well-known as high-dimensional, large-scale and deeply redundant[7,8]. In dynamic optimization Processing, high-dimensional data involves high computational cost[9]. Bellman mentions to the high-dimensional as the "curse of dimensionality"[10]. Microarray dataset is illustrative of this type of problem. a huge dataset with a small sample size with high dimensionality(number of features). The principal difficulty of microarray data analysis is to find a way for reducing both the dimension and redundancy of gene expression data through projection the data onto a more modest number of features which Only preserves the information of the small number of genes that required in cancer diagnosis as much as possible. This difficulty can be increased because of variability and noise. Noise in the dataset can be caused by deviations or errors whether provided in the data compilation phase through human error in transcribing information or due to conditions in the tolerance of the measurement equipment[11]. In general, noisy data tend to be affected by Machine learning algorithms, presenting it numerous difficult for learning algorithms to compose accurate models from the data. Therefore, to avoid unnecessary complexity in these models and improve the efficiency of the algorithm Noise should be reduced as much as possible. Several studies have shown the effect of noise, including (1)attribute noise refers to corruption in the values of one or more attributes erroneous attribute values, missing or unknown attributes values, and incomplete attributes or "to nor care " values and (2)class noise also referred as (label noise) is caused by contradictory models or misclassification [12].

Several types of research have shown that the accurate classification of distinct types of tumors is affected by irrelevant genes that measured in DNA microarray experiment[13]. To overcome this, two techniques plays as a crucial role to decrease the number of irrelevant genes so that the learning algorithm focuses only on useful training data for analysis and classify [14] (1)genes (features) selection and (2)genes (features) extraction.

Some cancer statistics provided by GLOBOCAN [15] showed that 14 million people suffered from cancer over the next two decades. According to the World Health Organization, 8 million people die from cancer and. It is expected that in 20 years their number will increase to 24 million [16]. Cancer is not the only disease. There are various types of cancer, many of which can be effectively treated today to eliminate, reduce or slow down the impact of the disease on the lives of patients. While a cancer diagnosis is made, some advances have been made in identifying genes related to the etiology of cancer.

The technology of DNA microarray provides a comprehensive source of gene expression data that are relevant to different fields including medicine especially cancer, and could provide an opening for the researchers to analyze genes profiles concurrently, which, against proper analysis, might enhance classified of patient gene expression profile, in which has becomes the most widely studied in biomedical research. Several microarray analyses have been Contributed significantly to examine the cancer genetic mechanisms, and utilized analytical approaches in order to classify different types of cancer or identify cancerous and noncancerous tissue. The real problem is managing microarray data with its high-dimension, Where classification algorithms become too complicated to understand the characteristics of gene expression. Due to the presence of more exceeding improper attributes in the dataset. Recently years, many researchers tried to analyze microarray data using machine learning techniques. Many approaches have been used to classify different types (include identify subtypes) of cancer such as, kNN k-Nearest Neighbors [17] , SVM Support Vector Machines[18],ANN Artificial Neural Networks[19], Bayesian Network, Decision Tree, NB Naïve Bayes[20], Genetic Algorithms, Rough Sets, Emerging Patterns, Self-Organizing Maps [20,21]. In addition to the distinction between cancerous and noncancerous Tumors that may progress aggressively. Several gene (feature) selection methods have been suggested to reduce the data dimensionality [22]. All these investigations are seeking to isolate the most significant genes from DNA microarray data in order to minimize the features space And therefore generate biologically meaningful interpretations of complex datasets.

978-1-7281-4813-7/19/$31.00©2019IEEE

This paper is distributed into four sections and the rest is structured as follows: Section II explains descriptions of DNA microarray technology. Section III discusses Challenges Analyzing Gene Expression microarray Data. Section IV includes feature selection methods (wrappers, filters, hybrid, ensemble and embedded) that applied on microarray cancer data.

II. DNA MICROARRAY

DNA microarray technology [23,24] is an evolving technology that has proposed an effective data collection method that can be used to measure the expression levels of thousands of genes simultaneously. Monitoring gene expression, which refers to the level of production of protein molecules determined by a gene, is one of the broad approaches used in genetics and molecular biology. Measuring mRNA instead of proteins is a standard method for measuring gene expression since the structure of the protein is different from the structure of the gene, so analyzing thousands of proteins will be difficult. In addition, mRNA sequences hybridize to their complementary RNA or DNA sequences, while proteins do not possess this property.

Experimental microarray steps include extracting mRNA from a tissue or cell sample. Then the mRNA is labeled using the reverse transcriptase enzyme (RT), which generates a complementary cDNA to the mRNA with fluorescent nucleotides, which ultimately leads to the labeling of the tumor and normal samples with various fluorescent dyes. Labeled cDNAs are then placed on the surface of DNA microarrays. Labeled cDNAs, which represent mRNA in a particular cell, will then bind to complementary base pairs at each point on the microarray, a process known as hybridization. Based on how the DNA binds together, a very active gene produces many RNA messengers, which means more labeled cDNAs that hybridize with the DNA, so a very bright fluorescent region will be generated. In addition, dimmer fluorescent spots are formed, representing slightly less active genes that produce less RNA messenger, so less cDNA is marked. Although fluorescence does not indicate that the gene is inactive, none of the RNA messengers was hybridized with DNA. Tumor samples and a normal sample compete with each other for synthetic complementary DNA on a microarray chip. Thus, each spot will be displayed in red, green or yellow (a combination of red and green) when scanning with a laser. A red spot indicates that this gene was more strongly expressed in cancer cells than in normal cells (increased regulation). A green spot indicates that this gene was severely repressed in normal cells (with down-regulation). If the spot turns yellow, this means that this gene was neither strongly expressed nor strongly repressed in cancer cells (equally expressed in normal and tumors) figure 1.

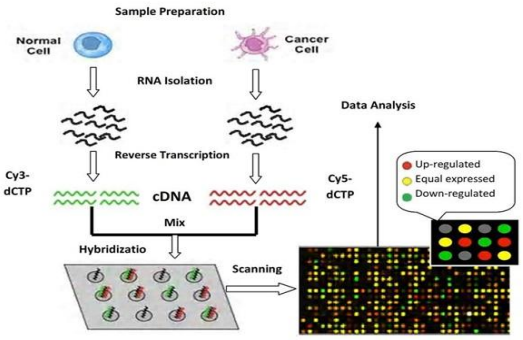

Fig. 1. Steps of a Microarray Experiment. (*figure caption*)

Finally, the gene expression data which produced by Microarray experiments are stocked as large matrix, where row represents the genes G {gene$_1$,…..,gene$_n$}, and the columns S {S_1,..., S_m} characterize the samples, and x_{ij} measured the expression intensity of the ith gene (i=1, …, n, n is the number of genes) in the jth sample (j=1, …, m, m characterizes the samples or the number of experiment conditions).

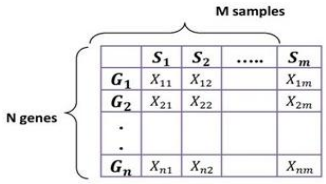

Fig. 2. The structure of a Microarray data matrix

Figure 2 display the structure of a DNA microarray matrix represented by N(dimension of genes) x M(sample or conditions involved in a particular Microarray experiment). Although the analysis of these dataset depends on the identification of research biological questions posed, analysis of microarray data poses new challenges, especially when gene expression data are usually of very high dimensions (contain a large number of N genes with the small number of M experiments). Moreover, the problem of classifying these data becomes much more challenging Due to the number of class labels exceeds five [25].

III. CHALLENGES ANALYZING GENE EXPRESSION MICROARRAY DATA

The analysis and interpretation of microarray data have faced many challenges need to be addressed. Some of these problems are as follows:

- Bias and confounding: which occurred when study, the design phase of microarray and can contribute to an incorrect conclusion. A systematic error in the design, recruitment, data collection or analysis that results in a confused estimation could perhaps cause bias. While confounding leads to a state of affairs in which the effect or association between the research variables of concern is distorted by the presence of another variable [26].
- For the analysis of expressions, a variety of platforms with microarrays are available, which differ in the variety of platform designs. As a result, cross-platform comparisons of gene expression studies are difficult. To overcome this problem, minimal information on the microarray experiment (MIAME) was developed [27] to improve reproducibility, sensitivity, and stability in the analysis of gene expression.
- Intrinsic characteristics of this microarray are large data (up to several tens of thousands of genes) with small sample sizes (usually less than 100), which cause significant problems, such as error estimation, which is strongly influenced by small samples [28], which in turn leads to misapplication of classification methods. In addition, inappropriate and noise genes, difficulty in constructing classifiers and multiple missing gene expression values As a result of multidimensional microarray data that represent microarray data, data overfitting suffers, which requires additional validation.
- Identify biologically significant changes that seek to generate biologically significant interpretations in gene expression is an integral criterion that should be

considered when analyzing microarray data, and not focus only on the accuracy of cancer classification. The disclosure of biological information on the cancer classification process can help experts in the development and planning of more appropriate treatments for cancer patients.

- Mislabeled samples which are usually resulted by the similarity of different subtype of disease[29], could decrease classification accuracy seriously, thus led to the inaccurate conclusion about gene expression patterns.

IV. Features Selection

Feature selection is the process of decreasing the dimensionality of the Microarray data with the aim of developing the classification accuracy that requires identifying a subset of highly distinctive genes [30] that best differentiate biological samples of distinct types. However, due to the size of the data to be processed that exponentially increased beside that the most of the genes measured in DNA microarray experiments are not relevant to improve classification accuracy or are redundant[31] or not be related to the cancer, feature selection has shown a requirement before obtaining any kind of classification. The implementation of the feature selection process is not only due to the need for the elimination of features that are not relevant or redundant or non-distinctive but also to more accurately correlate gene expression to cancer diseases. The objectives of feature selection techniques are many, the principal ones are to avoid overfitting the data, in which make further analysis possible, besides To provide faster and more extra cost-effective models, and in order to improve classifier prediction performance. In feature selection, the number of the genes measured in DNA microarray experiments is reduced according to some criterion such as eliminate irrelevant or redundant genes and only selecting genes that have a high level of activity.

In the context of classification, the feature selection technique typically broadly classified into three machine learning families:

1) The supervised gene selection method selects discriminative features according to importance and relevance with regard to the class. Some of the works of literature of supervised gene selection techniques are discussed in [37].

2) Unsupervised feature selection method, there is no prior between the class, it estimates feature relevance by taking advantage of innate structures of the data, such as data distribution, variation, and separability.

3) A Semi-supervised feature selection integrates labeled and unlabeled data as additional information for finding the discriminating features, and to improve an unsupervised feature selection performance. Most of the semi-supervised algorithms construct the similarity matrix and then rely on it to selecting the features.

According to [38], the feature subsets classify into four categories: (a) completely irrelevant and noisy features, (b) weakly relevant and redundant features, (c) weakly relevant and non-redundant features, and (d)strongly relevant features. An optimal subset contains the categories (c) and (d). Strongly relevant features are features contains high necessary information which has the principal role for enhancement of judicial power and prediction accuracy in the target classification. Redundant features typically contain worthless information and which not provide any data about the discriminating features, but they may have correlated with relevant features. Weakly relevant features may be valuable in improving prediction accuracy, unlike irrelevant feature which doesn't offer any improvement, so it is should be eliminated in order to establish a good model prediction. Thus, all strongly relevant features and some weakly relevant features should be beheld as a good subset, and irrelevant, redundant or noisy features should be extinguished. As a termination, the process of Feature selection typically divided into five categories: filter methods select the features independently without using any classification algorithm, whereas wrapper methods utilize the learning techniques to evaluate the quality of the selected subset. Embedded methods which combine the space of feature subsets and hypotheses, and the classifier construction[32,33]. The hybrid method may be either made by integrating two feature selection approaches, or two different methods of the same criterion. The most common hybrid method combined wrapper and filter methods[39]. Finally, the ensemble method aims to introduce variety and increase the regularity of the feature selection process, by overcoming the perturbation and instability problems in many features selection algorithms [40].

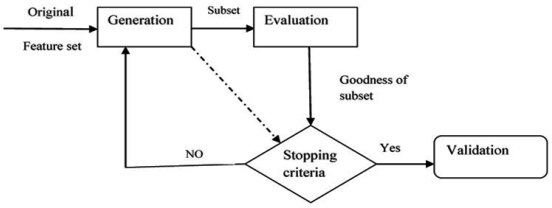

Fig. 3. The feature selection process with validation[71]

A. Filter Method

Filters, known as the open-loop method, select elements independently, without regard to the classification algorithm. This makes them very computationally efficient because it has a very simple calculation method and can simply be scaled to large-scale microarray datasets since it only has a short lead time. In general, filtering methods evaluate the relevance of features by measuring the statistical or geometric properties of the subset to be evaluated based on four types of evaluation criteria, namely, dependency, consistency, distance, and information [34]. In most cases, he uses gene ranking methods as standard statistical criteria for ordering and selecting traits. An assessment of the relevance of an object is calculated using various statistical and multidimensional methods, and based on the threshold value, only objects with the highest rating are selected from the large set, and other objects with a low rating are excluded.

The filter method evaluates the characteristics without using any classification algorithm, so it can provide general solutions for various classifiers. In addition, the bias in the selection of attributes does not correlate with the bias in the learning algorithm, therefore, it has the best generalization property [35]. The filtering method is usually divided into two stages. The first stage involves ranking the object in accordance with specific assessment criteria, it can be either a multivariate method or a one-dimensional method. The multidimensional method is able to process duplicated and redundant objects, in addition to finding relationships between objects, while one-dimensional methods examine each object separately and they are ranked independently in the space of objects. At the second stage, higher-ranking functions are selected to stimulate machine learning classification models. The following methods have been proposed for ranking genes in a data set according to their significance [36]:

- (Univariate) Unconditional modeling of a mixture involves two different biological statuses of gene activity (on or off) and using the probability of overlap of the mixture to expect if the binary state of the gene will affect the classification or not.

- (Univariate) The information benefit ranking used to find a good approximation of the conditional distribution, P (C | F), where F is the common feature vector and C is the class label. Obtaining information is one of the most common methods for evaluating attributes used in classification systems as a replacement for conditional distribution.

- Markov's (multivariate) block filtering removes functions that are conditionally independent of the class label without compromising the accuracy of class prediction.

- (Multivariate) Correlation-based feature selection (CFS) is a simple filtering algorithm that ranks attributes according to a correlation-based heuristic evaluation function [41]. The function evaluates subsets made from attribute vectors that are closely related to the class label but not related to each other. The CFS method considers that feature that shows low correlation with the class are irrelevant and should, therefore, be ignored. On the other hand, redundant elements should be studied, since they will be closely related to one or more residual elements. The acceptance of a function will depend on the range in which it predicts classes in regions of instance space not yet predicted by other functions. Alomari and Alzboon used The Correlation based Feature Selection (CFS) with Greedy Stepwise search method is proposed for genes selection Also, applied multiple classifiers as Decision Table, JRip, and OneR applied on the original datasets to show the quality of each of them. First, the all microarray datasets were filtered using Correlation-based Feature Selection (CFS) algorithm. then, the filtered datasets were tested against the applied classifiers. This was done in order to compare the classification accuracy of the dataset with the one before filtration. The comparative analysis proved that the accuracy of all classifiers is improved using filtered datasets compared with their accuracy on the original datasets. This indicates that the feature selection by CFS not only improved the efficiency of the classification process but also its accuracy is enhanced as seen in table I that show the Average accuracy for the microarray data

- (multivariate) fast correlation filter (FCBF) [42] - a simple algorithm that works with large-dimensional data and is effective for eliminating both redundant and irrelevant genes (functions). By measuring two types of correlation, an object class, and a feature. It works by choosing a subset of functions that are closely related to the class. After that, he applies three heuristics to eliminate unnecessary or unnecessary functions and leave those functions that are more relevant to the class. However, it does not take into account the interaction between functions.

- (Multivariate) Minimum Redundancy Maximum Relevance (mRMR) [43]: the method selects the objects that have the most relevance with a class label while minimizing redundancy in each class using a variety of statistical indicators. Both optimization criteria (minimum redundancy, maximum relevance) are based on mutual information (MI), which measures the information that a random variable can give about the other, especially the class label and gene activity.

- (Multivariate) ReliefF [46], stretching the original Relief, is an algorithm that selects the most distinguishable features between different classes. It works by randomly fetching an instance (sample) from the data, and then finds its nearest neighbor from the same and opposite class, it gives, based on its neighbors, more weight to functions that help distinguish it from neighbors of another class [44, 45]. The relief method can be used in all situations, has low bias, includes the interaction between functions, in addition to being able to interact with multiclass problems and more reliable processing with noisy data.

- The INTERACT filter method [47] consists of two main steps. At the first stage, objects are ranked in descending order based on their SU values. At the second stage, the functions are evaluated in turn, starting from the end of the list of rank objects. Elements that are less than the set threshold are deleted, and objects remain selected. To do this, it uses the same FCBF filter measures, in addition to the consistency contribution, which is an indicator of how deleting a function will affect the consistency.

TABLE I. AVERAGE ACCURACY FOR THE 8 MICROARRAY DATA [48]

classifier	Accuracy for the full training and cross validation method								Average
	Breast	CNS	Colon	Leukemia	Lung	MII	Ovarian	SRBCT	
Decision Table	61.0	61.0	74.4	85.0	83.7	83.9	96.8	72.2	77.3
Decision Table-CFS	69.7	73.3	82.9	86.8	86.5	86.5	96.9	73.9	82.1
JRip	59.8	60.0	74.4	86.0	88.3	82.2	97.2	85.7	79.2
Jrip-CFS	69.1	71.5	82.4	88.1	91.6	87.1	97.3	89.5	84.6
OneR	53.7	57.8	71.0	85.0	75.3	78.5	96.9	58.3	72.0
OneR-CFS	59.5	66.0	77.1	85.0	76.1	79.0	96.9	56.7	74.6

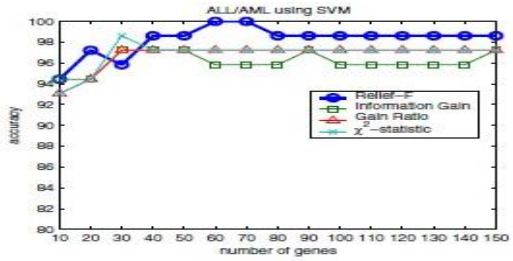

Fig. 4. Comparison between ReliefF, Information Gain, Information Gain Ratio, and test on ALL / MLL Leukaemia datasets[45]

Although all Filter methods are computationally exceeding efficient, fast, and scalable, and provides more generalized result and may be more accurate in classifying information, the disadvantage of the above methods is that they evaluate only the relevance between the feature without taking into account the dependency between them. Furthermore, none of the above methods have indicated whether the results are biologically relevant or not so the genes identified by them do not provide any proven biological significance.

B. Wrapper Method

While the Filter method selects the elements independently, not taking into account the hypothesis of the model, thus, the

main disadvantage of the filtering method is that it completely ignores the importance of functions with respect to the induction algorithm, since some functions will be based on offsets and heuristic algorithm induction. On the other hand, the wrapper method tends to work better when selecting objects; they include the method of selecting objects in the classification algorithm. This leads to a great shortage of packers due to the growth of its computational costs as the exponential growth of the feature space. Inefficiency in computing becomes an important aspect when tens of thousands of functions are considered. In addition, there is a risk of re-equipment in accordance with the small sample size of the DNA microarray data. As a result, the wrapper approach did not receive the same attention as filtering methods, and it was avoided in the literature. Wrapping methods [51] provide a simple and reliable way to overcome the problem of selecting traits by searching for a subset of genes in the first step. While the second step evaluates the quality of each obtained subset of genes by assessing the percentage of accuracy of the particular classifier used, training the classifier only with the genes found. Then repeat the first and second steps until you find the best quality subgroup of genes.

The wrapper methods utilize the induction algorithm as a black-box manner. The set of features is generated by the feature search component, then the feature evaluation component utilizes classification error rate or performance accuracy as a feature evaluation criterion. The feature set with highest classifier accuracy is selected as the most discriminative subset. Because the search space of the 'n' genes (features) is the O(2n), this exhaustive search is impractical, particularly when we are dealing with high-dimensional microarray data, unless the size of 'n' is small, i.e., the NP-hard(nondeterministic polynomial time) problem. To overcome these issues, different search strategies have been used to find the best subset such as branch and bound method, Genetic algorithm (GA), particle swarm optimization (PSO). Wrapper methods are separated into 2 categories: deterministic and randomized search algorithm.

1) Deterministic Wrapper method: Uses a greedy strategy to select features based on a local change. Although there are many alternatives to this straightforward method, the creation of the subset is basically incremental[56]. Sequential backward selection(SBS), sequential forward selection, sequential floating forward selection, and sequential floating backward selection, are examples of this procedure. A combination of a wrapper and sequential forward selection (SFS) has been used to examine breast cancer. SFS is a deterministic feature selection method using a hill-climbing search strategy to add all possible single-attribute expansions to the current subset and evaluate them. SFS starts with an empty subset of genes and genes are selected sequentially, one at a time, until no further improvement is achieved in the evaluation function. The feature that leads to the best score is added permanently [52,53]. With regard to classification, K-nearest neighbors, support vector machines (SVMs), and probabilistic neural networks were used in an attempt to classify between cancerous and noncancerous breast tumors [54]. perfect results were achieved using SVMs. Two methods based on SVMs are very widely used to classify DNA microarray cancer datasets:

- Gradient-based-leave-one-out gene selection (GLGS) [55,56] was originally introduced to select parameters for SVMs by applying PCA to a Microarray cancer dataset. Then, a new low-dimensional space was calculated and optimized using a gradient-based algorithm. After that, the pseudo scaling coefficients of the original genes are calculated. Finally, gene selection is sequentially based on a correlation coefficient.

- Leave-one-out calculation sequential forward selection (LOOCSFS) is based on sequential direct selection (SFS) and is very widely used to select signs of cancer. It works by adding genes in an initially empty set and then calculates the cross-validation error with the list omitted [58]. This is basically an objective assessment of generalization errors using SVM and C Bound. C Bound (decision boundary) is used as an additional criterion when different features in a subset have the same cross-validation error without residue (LOOCVE) [59, 60, 61]. SFS can be used together with the recursive reference vector algorithm (R -CVM), which selects important genes or biomarkers [62]. In addition, SFS may also add some restrictions [63] on the size of the subset to be selected. Based on the minimum error of the SVM support vector machine, the contribution factor is ranked and calculated for each gene.

2) Randomized Wrapper Method: Genetic Algorithm (GA) is a randomized search and optimization algorithm that mimics evolution and natural genetics. It was used in [64,65] for recognition of cancer of binary and multiclass cancer. The shell, called the Best Incremental Ranked Subset (BIRS) [66], is an algorithm presented for gene selection. He evaluates genes based on their value and class label and then uses incremental ranked utility (based on Markov form) to identify excess genes. Linear discriminant analysis was used in combination with genetic algorithms [67].

C. Embedded Method

Although the filter method requires less time, one of the main limitations of the filter approach is that it does not depend on the classifier, which usually leads to poor performance, in contrast to wrapper methods. However, the wrapper approach is computationally expensive, which is especially compounded by the high dimensionality of microarray cancer data sets. In this case, a compromise for researchers is to use hybrid or built-in approaches that depend on the classifier to establish criteria for ranking objects by including the selection of features in the educational process, which reduces the computational cost due to the classification process required for each subset. Probably the best-known built-in method is Support Vector Machine based on Recursive Feature Elimination (SVM-RFE) [68].

D. Hybrid Method

The hybrid combines two different methods, filtering and filling, and uses these two methods to achieve the best performance, trying to inherit the advantages of both methods by combining their complementary advantages [69]. The combination can be either a filter-filter, a filter-wrapper or a filter-filter-wrapper, where a subset of the genes obtained by one method is supplied as input to another selection algorithm. As a rule, the choice of the initial subgroup of genes is carried out using the filter method, which also helps to remove excess genes. The selection of the initial set of functions can also be made by applying vertically any combination of several filtering methods. Then the wrapper method selects the optimal object from the objects that were selected in the previous step. The ability of this method to increase the efficiency and accuracy of

forecasting is due to the use of various assessment criteria. The most common hybrid method is mentioned in [70].

E. Ensemble Method

The purpose of the ensemble method, in addition to solving the problems of instability and turbulence in many feature selection algorithms, is to add variety and increase the regularity of the feature selection process. This method combines relevant features selected by various classification methods (ranking approaches), which leads to the selection of a subset of features and is either aggregated or intersected to obtain the most relevant subset of features. In general, ensemble learning was applied to the classification and was recently applied to the selection of microarray genes, and the method was robust and stable when working with them [71].

TABLE II. SUMMARIZATION OF FEATURE SELECTION TECHNIQUES FOR CANCER CLASSIFICATION

Source	Dataset	FS	Used algorithms	Efficancy AND ACCURACY of algorithm
Guo& Shun[2016]	Tumors-11 and Colon datasets Breast- Colon	**Filter** Feature Selection Method	Regularized Logistic Regression (RLR)	is effective and competitive compared with(MSVM-RFE- F-test- LLFS-KernelPLS)
[Chen, Huihui 2016]	Lung Prostate, brain_Tumor, ALLAML Lymphoma		SVM, k-nearest neighbor KBCGS algorithm	KBCGS method (1) fast and efficient; (2) model-free and parameter-free; (3) easy to extend
[Hoque,2016]	Breast Cancer,Leukemia,colon		(FMIFS-ND)algorithm KNN, KNN-ND, NaiveByes, Decision tree , Random forest	(FMIFS-ND) method give high and effective accuracy and has good classification result
[Wang, Shuqin, 2017]	lung, leukemia, breast		(CAM) algorithm, SVM, KNN, CFS, IG, Relief, Naive Bayes (NB)	CAM is effective and attractive way to feature selection
[Mishra, Debahuti ,2011]	Leukemia		SVM, k-means clustering, and SNR, SNR, (kNN), (PNN), (fNN)	k-means clustering and SNR for selecting differentially expressed genes gives better result for SVM
[Pour, Ali Foroughi ,2014]	--		IA and 2MNC algorithms	IA and 2MNC appear very robust to violations on the assumed model and excellent performance
[Yang, Pengyi,2010]	Leukemia, Colon, Liver, MLL		MFGE, hybrid with GE, GA/KNN, and Gain Ratio filter algorithm	MF-GE system has a higher average classification accuracy for all datasets
[Peng, Hanchuan, 2005]	HDR-ARR-NCI-lymphoma		(minimal-redundancy-maximalrelevance mRMR)	The classification accuracy can be significantly improved based on mRMR feature selection.
[GÜÇKIRAN, K., 2019]	Breast ,Leukemia,Lung ,ALL,COLON		SVM, MLP, LASSO, Relief	MLP accuracy heavily depends on the initial values. LASSO can further decrease coefficient to zero using L1 regularization.
[Al-Batah, M., 2019]	Breast ,,Leukemia,Lung ,Ovarian		(CFS) algorithm,Decision Table, JRip, and OneR	average accuracy of JRip was better than Decision Table and OneR
[Ooi, C. H ,2003]	NCI60	**Wrapper FS method**	genetic algorithms (GAs) MLHD Classifier GA/MLHD	GA-based approach are that it automatically determines the optimal predictor set size and the delivery of predictive accuracies
[Gutlein,2009]	Leukemia Lung- MLL-Leukemia Prostate		Sequential Forward Selection (SFS). ORDERED-FS, naive Bayes (NB)	Sequential Forward Selection (SFS)produce smaller subsets without marked changes in accuracy

TABLE III. ADVANTAGES AND DISADVANTAGES OF WRAPPER, FILTER, ENSEMBLE, EMBEDDED, HYBRID METHODS

Method	Advantages	Disadvantage
Filter	Fast, scalable , Independent of the classifier, Faster than the wrapper method, Better computation complexity than the wrapper method	Ignores feature dependencies , Ignores interaction with the classifier, Redundant features may be included

Method	Advantages	Disadvantage
Wrapper	Higher performance accuracy than filter, Consider the dependence among features, Interacts with the classifier, Prone to local optima	Classifier dependent selection, Computationally intensive, Higher risk of over-fitting
Embedded	Higher performance accuracy than the filter, Interacts with the classifier, Prone to local optima, Less prone to over-fitting than the wrapper, Better computational complexity than the wrapper	Classifier dependent selection, Computationally intensive, Consider the dependence among features
Hybrid	Higher performance accuracy than the filter, Better computational complexity than the wrapper, More flexible and robust upon high dimensional data, Less prone to over-fitting than the wrapper.	Classifier specific, Difficult to understand an ensemble of classifiers.
Ensemble	Less prone to over-fitting, More scalable for high dimensional datasets.	Understanding an ensemble of classifiers is difficult.

V. CONCLUSION

This paper discusses various methods for reducing the size of data on multidimensional microarray cancer, which are necessary to obtain meaningful results when increasing the amount of data to be analyzed. Various methods for feature selection for these microarrays have been described, as well as their advantages and disadvantages.

REFERENCES

[1] Leung, Yuk Fai, and Duccio Cavalieri. "Fundamentals of cDNA microarray data analysis." TRENDS in Genetics 19.11 (2003): 649-659.

[2] Bolón-Canedo, Verónica, et al. "A review of microarray datasets and applied feature selection methods." *Information Sciences* 282 (2014): 111-135.

[3] Salem, Hanaa, Gamal Attiya, and Nawal El-Fishawy. "Classification of human cancer diseases by gene expression profiles." Applied Soft Computing 50 (2017): 124-134.

[4] Lv, Jia, et al. "A multi-objective heuristic algorithm for gene expression microarray data classification." Expert Systems with Applications 59 (2016): 13-19.

[5] Lu, Huijuan, et al. "A hybrid feature selection algorithm for gene expression data classification." Neurocomputing 256 (2017): 56-62.

[6] Attiya, G., El-Fishawy, N., 2017. Classification of human cancer diseases by gene expression profiles. Appl. Soft. Comput. 50, 124-134.

[7] Brazma, Alvis, and Jaak Vilo. "Gene expression data analysis." FEBS Letters 480.1 (2000): 17-24.

[8] Sherlock, Gavin. "Analysis of large-scale gene expression data." Current opinion in immunology 12.2 (2000): 201-205.

[9] Kung, Sun-Yuan, and Man-Wai Mak. "Feature selection for genomic and proteomic data mining." Machine Learning in Bioinformatics 4 (2009): 1.

[10] Jain, Anil, and Douglas Zongker. "Feature selection: Evaluation, application, and small sample performance." IEEE transactions on pattern analysis and machine intelligence 19.2 (1997): 153-158.

[11] Nettleton, David F., Albert Orriols-Puig, and Albert Fornells. "A study of the effect of different types of noise on the precision of supervised learning techniques." Artificial intelligence review33.4 (2010): 275-306.

[12] Zhu, Xingquan, and Xindong Wu. "Class noise vs. attribute noise: A quantitative study." Artificial intelligence review 22.3 (2004): 177-210.

[13] Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." science 286.5439 (1999): 531-537.

[14] C. Shang and Q.shen, "Aiding classification of gene expression data with feature selection: A comparative study", Vol1, 2006, pp 68-76

[15] Plummer, Martyn, et al. "Global burden of cancers attributable to infections in 2012: a synthetic analysis." The Lancet Global Health 4.9 (2016): e609-e616.

[16] Ferlay, J., Soerjomataram, I., Dikshit, R. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. Int J Cancer. 2015; 136(5): E359–86.

[17] Dudoit, Sandrine, Jane Fridlyand, and Terence P. Speed. "Comparison of discrimination methods for the classification of tumors using gene expression data." Journal of the American statistical association 97.457 (2002): 77-87.

[18] Brown, Michael PS, et al. "Knowledge-based analysis of microarray gene expression data by using support vector machines." Proceedings of the National Academy of Sciences97.1 (2000): 262-267.

[19] Khan, Javed, et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." Nature medicine 7.6 (2001): 673.

[20] Su, Yang, et al. "RankGene: identification of diagnostic genes based on expression data." Bioinformatics 19.12 (2003): 1578-1579.

[21] Peng, Yonghong. "A novel ensemble machine learning for robust microarray data classification." Computers in Biology and Medicine 36.6 (2006): 553-573.

[22] Mohan, Abhilash, et al. "Automatic classification of protein structures using physicochemical parameters." Interdisciplinary Sciences: Computational Life Sciences 6.3 (2014): 176-186.

[23] Chee, Mark, et al. "Accessing genetic information with high-density DNA arrays." Science 274.5287 (1996): 610-614.

[24] Fodor, Stephen P., et al. "Light-directed, spatially addressable parallel chemical synthesis." science 251.4995 (1991): 767-773.

[25] Ooi, C. H., and Patrick Tan. "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data." *Bioinformatics* 19.1 (2003): 37-44.

[26] Tinker, Anna V., Alex Boussioutas, and David DL Bowtell. "The challenges of gene expression microarrays for the study of human cancer." *Cancer cell* 9.5 (2006): 333-339.

[27] Brazma, Alvis, et al. "Minimum information about a microarray experiment (MIAME)—toward standards for microarray data." *Nature Genetics* 29.4 (2001): 365.

[28] Dougherty, Edward R. "Small sample issues for microarray-based classification." *Comparative and functional genomics* 2.1 (2001): 28-34.

[29] Khan, Javed, et al. "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks." *Nature medicine* 7.6 (2001): 673.

[30] Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics." *bioinformatics*23.19 (2007): 2507-2517.

[31] Golub, Todd R., et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring." *science* 286.5439 (1999): 531-537.

[32] Chandrashekar, Girish, and Ferat Sahin. "A survey on feature selection methods." *Computers & Electrical Engineering* 40.1 (2014): 16-28.

[33] Miao, Jianyu, and Lingfeng Niu. "A survey on feature selection." *Procedia Computer Science* 91 (2016): 919-926.

[34] Dash, Manoranjan, and Huan Liu. "Feature selection for classification." *Intelligent data analysis* 1.1-4 (1997): 131-156.

[35] Ding, Chris, and Hanchuan Peng. "Minimum redundancy feature selection from microarray gene expression data." *Journal of bioinformatics and computational biology* 3.02 (2005): 185-205.

[36] Xing, Eric P., Michael I. Jordan, and Richard M. Karp. "Feature selection for high-dimensional genomic microarray data." *ICML*. Vol. 1. 2001.

[37] Ang, Jun Chin, et al. "Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection." IEEE/ACM transactions on computational biology and bioinformatics 13.5 (2015): 971-989.

[38] Yu, Lei, and Huan Liu. "Efficient feature selection via analysis of relevance and redundancy." Journal of machine learning research 5.Oct (2004): 1205-1224.

[39] Peng, Yonghong, Zhiqing Wu, and Jianmin Jiang. "A novel feature selection approach for biomedical data classification." Journal of Biomedical Informatics 43.1 (2010): 15-23.

[40] Awada, Wael, et al. "A review of the stability of feature selection techniques for bioinformatics data." 2012 IEEE 13th International Conference on Information Reuse & Integration (IRI). IEEE, 2012.

[41] Hall, Mark A. "Correlation-based feature selection of discrete and numeric class machine learning." (2000).

[42] Yu, Lei, and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution." Proceedings of the 20th international conference on machine learning (ICML-03). 2003.

[43] Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy." IEEE Transactions on Pattern Analysis & Machine Intelligence 8 (2005): 1226-1238.

[44] Mercier, G., et al. "Biological detection of low radiation doses by combining results of two microarray analysis methods." Nucleic Acids Research 32.1 (2004): e12-e12.

[45] Wang, Yuhang, and Fillia Makedon. "Application of Relief-F feature filtering algorithm to selecting informative genes for cancer classification using microarray data." Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004. IEEE, 2004.

[46] Hall, Mark A., and Lloyd A. Smith. "Practical feature subset selection for machine learning." (1998): 181-191..

[47] Zhao, Zheng, and Huan Liu. "Searching for interacting features in subset selection." Intelligent Data Analysis 13.2 (2009): 207-228.

[48] Al-Batah, Mohammad, et al. "Gene Microarray Cancer Classification using Correlation Based Feature Selection Algorithm and Rules Classifiers." International Journal of Online Engineering 15.8 (2019).

[49] GÜÇKIRAN1CANTÜRK and ÖZYILMAZ " DNA Microarray Gene Expression Data Classification Using SVM, MLP, and RF with Feature Selection Methods Relief and LASSO." Journal of Natural and Applied Sciences Volume 23, Issue 1, 126-132, 2019..

[50] Chandrashekar, Girish, and Ferat Sahin. "A survey on feature selection methods." Computers & Electrical Engineering 40.1 (2014): 16-28.

[51] Kohavi, Ron, and George H. John. "Wrappers for feature subset selection." Artificial intelligence 97.1-2 (1997): 273-324.

[52] Hira, Zena M., and Duncan F. Gillies. "A review of feature selection and feature extraction methods applied to microarray data." Advances in bioinformatics 2015 (2015).

[53] Pudil, Pavel, Jana Novovičová, and Josef Kittler. "Floating search methods in feature selection." Pattern recognition letters15.11 (1994): 1119-1125.

[54] Osareh, Alireza, and Bita Shadgar. "Machine learning techniques to diagnose breast cancer." 2010 5th International Symposium on Health Informatics and Bioinformatics. IEEE, 2010.

[55] Chapelle, Olivier, et al. "Choosing multiple parameters for support vector machines." Machine learning 46.1-3 (2002): 131-159.

[56] Xia, Xiao-Lei, Huanlai Xing, and Xueqin Liu. "Analyzing kernel matrices for the identification of differentially expressed genes." PloS one 8.12 (2013): e81683.

[57] Ruiz, Roberto, José C. Riquelme, and Jesús S. Aguilar-Ruiz. "Incremental wrapper-based gene selection from microarray data for cancer classification." Pattern Recognition 39.12 (2006): 2383-2392.

[58] Ambroise, Christophe, and Geoffrey J. McLachlan. "Selection bias in gene extraction on the basis of microarray gene-expression data." Proceedings of the national academy of sciences 99.10 (2002): 6562-6566.

[59] Liu, Qingzhong, et al. "Feature selection and classification of MAQC-II breast cancer and multiple myeloma microarray gene expression data." PloS one 4.12 (2009): e8250.

[60] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3 (2002): 389-422.

[61] Liu, Qingzhong, et al. "Gene selection and classification for cancer microarray data based on machine learning and similarity measures." BMC Genomics 12.5 (2011): S1.

[62] Jirapech-Umpai, Thanyaluk, and Stuart Aitken. "Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes." BMC Bioinformatics 6.1 (2005): 148.

[63] Gutlein, Martin, et al. "Large-scale attribute selection using wrappers." 2009 IEEE symposium on computational intelligence and data mining. IEEE, 2009.

[64] Liu, Jane Jijun, et al. "Multiclass cancer classification and biomarker discovery using GA-based algorithms." Bioinformatics21.11 (2005): 2691-2697.

[65] Li, Leping, et al. "Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method." Combinatorial chemistry & high throughput screening 4.8 (2001): 727-739.

[66] Ruiz, Roberto, José C. Riquelme, and Jesús S. Aguilar-Ruiz. "Incremental wrapper-based gene selection from microarray data for cancer classification." Pattern Recognition 39.12 (2006): 2383-2392.

[67] Huerta, Edmundo Bonilla, Béatrice Duval, and Jin-Kao Hao. "Gene selection for microarray data by an LDA-based genetic algorithm." IAPR international conference on pattern recognition in bioinformatics. Springer, Berlin, Heidelberg, 2008.

[68] Guyon, Isabelle, et al. "Gene selection for cancer classification using support vector machines." Machine learning 46.1-3 (2002): 389-422.

[69] Kabir, Md Monirul, Md Monirul Islam, and Kazuyuki Murase. "A new wrapper feature selection approach using neural network." Neurocomputing 73.16-18 (2010): 3273-3283.

[70] Apolloni, Javier, Guillermo Leguizamón, and Enrique Alba. "Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments." Applied Soft Computing38 (2016): 922-932.

[71] Shen, Qiang, Ren Diao, and Pan Su. "Feature Selection Ensemble." Turing-100 10 (2012): 289-306.

[72] Guo, Shun, et al. "A centroid-based gene selection method for microarray data classification." Journal of theoretical biology 400 (2016): 32-41.

[73] Chen, Huihui, Yusen Zhang, and Ivan Gutman. "A kernel-based clustering method for gene selection with gene expression data." Journal of Biomedical Informatics 62 (2016): 12-20.

[74] Hoque, N., et al. "A fuzzy mutual information-based feature selection method for classification." Fuzzy Information and Engineering 8.3 (2016): 355-384.

[75] Wang, Shuqin, and Jinmao Wei. "Feature selection based on measurement of ability to classify subproblems." Neurocomputing 224 (2017): 155-165.

[76] Mishra, Debahuti, and Barnali Sahu. "A signal-to-noise classification model for identification of differentially expressed genes from gene expression data." 2011 3rd International Conference on Electronics Computer Technology. Vol. 2. IEEE, 2011.

[77] Pour, Ali Foroughi, and Lori A. Dalton. "Optimal Bayesian feature selection on high dimensional gene expression data." 2014 IEEE Global Conference on Signal and Information Processing (Globalsip). IEEE, 2014.

[78] Yang, Pengyi, et al. "A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data." BMC Bioinformatics 11.1 (2010): S5.

[79] Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy." IEEE Transactions on Pattern Analysis & Machine Intelligence 8 (2005): 1226-1238.