ISSN 2278-3091

Volume 9, No.5, September - October 2020

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse275952020.pdf

https://doi.org/10.30534/ijatcse/2020/275952020

Hybrid feature selection methods for the Classification of Cancer in Micro-array Gene expression data: a Survey

Peddarapu Rama Krishna[1], Dr.Pothuraju Rajarajeswari[2],

1 Research Scholar, Department of Computer Science and Engineering, KoneruLakshmaiah Education Foundation, Guntur, Andhra Pradesh,India, Email: peddarapuramakrishna@gmail.com

2 Professor, Department of Computer Science and Engineering, KoneruLakshmaiah Education Foundation, Guntur, Andhra Pradesh,India, Email: rajilikhitha@gmail.com

ABSTRACT

For the study of disease diagnostics, microarray technology is widely used with gene expression rates. Researchers simultaneously studied the degree of expression of thousands of genes through the advent of DNA microarray technology. The Micro-array data analysis is the method to remove redundant and obsolete genes, to identify the most significant genes. Cancer identification is one of the most significant applications of the study of micro-array results. The efficiency of microarray technology depends on measurement precision, use of data processing techniques, research methods and statistical modeling. But still the curse of dimensionality and the curse of sparseness are a challenge to classify the gene expression profile. A collection of features (genes) is one of the most successful approaches to face these challenges. The function selection methods are used by eliminating noisy and non-relevant features which develop classification performance to obtain an informative representation. Within the literature, there are numerous works to pick the main features from the micro-array. This paper explores the new methods for selecting hybrid characteristics for selecting major genes from the findings of micro-array expression for the diagnosis of cancer.

Key words: Micro-array, Gene expression, Cancer classification, Bio-inspired, Gene selection, Feature selection, Hybrid approach.

1.INTRODUCTION

In the fields of bio-informatics and biotechnology [1, 2] new DNA-focused technologies have been used. A typical microarray technology calls for DNA template hybridization of the mRNA molecules. It results in the development of techniques of gene selection, namely, the discovery of organic patterns which are monitored and unattended [3]. Cancer diagnosis is one of Microarray's most significant data analysis applications. The cancer-pathology research is a study of the cancer-causing genes, specifically the gene responsible for mutation of the disease. This means that different genes vary in level of expression. Nevertheless, the classification of the gene expression profile is a challenge and a (NP)-hard problem. [1]

Not every gene contributes to cancer. Most genes are unrelated to or negligible for clinical diagnosis. Microarray Data Analysis has two major issues. Two main problems apply to microarray data analyses: First, a high-dimensional microarray data set includes a few thousand genes, i.e. a Small sample number, typically ten samples, and has low data sparsity. Second, gene expression data are highly complex; genes are directly or indirectly related. Standard machine learning approaches have failed because they are more suitable if more examples than features are available.

Algorithms for reducing the dimension or selecting features (gene) were added to try to solve these problems. Gene selection strategies are classified into three categories: filter, wrapper and embedded. With its wide-ranging statistical properties, the filter methodology provides an individual evaluation. The wrapper approach uses learning strategies to pick the best subset of characteristics. A hybrid approach is developed for filtering and wrapping [2]. This combines a filter approach with a high-performance wrapper approach. The hybrid approach focuses on two steps. Firstly, a preprocessing step for filtering noise off and secondly, wrapping techniques which use ideal functions to suit the subset. The success of this approach depends on two factors: the correct taxonomy and the carefully chosen genes. The study is structured as a wrapper tool for evaluating and comparing current hybrid approaches by means of bio-inspired evolutionary methods.

A remarkable approach to reduction of data size (i.e. leukemia and colon) is the selection of functions [6]. One of the key problems was investigated as a "curse of dimensionality" by researchers in the DNA microarray. The design of features [or functional extraction] is primarily correlated with the collection of elements, which may also reduce complex dimensionality [8, 9]. For researchers working in machine learning, microarray data poses a significant challenge. Due to the chance that many fields have been found in contrast with very few samples, there is a high probability that "false positive" are detected. Due to the possibility that there are too many groups compared to so few samples, there is a high chance that 'false positive' are found. Robust procedures are required for validating models and evaluating their probability.

2.BACKGROUND

In this segment we shall discuss basic microarray technology concepts. Next, a summary of Microarray technology and the gene expression profile for Micro-array will also be presented. Instead we will explore the meaning of the data and its types.

2.1 The Gene Expression Profile for Microarray data: DNA microarray technology has become a powerful means of monitoring the gene expressiveness of an organism for biologists [3]. Generally, data on gene expression include thousands of genes and a limited number of specimens (high dimensionality). This also has other features that are outdated and redundant. Scientists are assisted in calculating the expression rates of several genes simultaneously. It is often used by medical professionals to understand the causes and treatments of disease [4]. All diseases have not been caused by a single form of mutation. The data technology for micro-array gene expression has had a significant effect on cancer care [5]. This has been used extensively to classify genes associated with cancer using methods of selection [6].

2.2 Data Analysis: The data analysis of gene expression is the latest gene discovery approach whereas the old and obsolete genes[7]. The analysis of data is an informative tool. Three types of microarray data analysis, namely class comparison, prediction of the class, and class discovery [8], are currently available. Community recognition is sometimes referred to as gene discovery, as Fig 1. The variations between function selections methods are a) filter, b) mask, c) embedded. [56]. Also known as classification, class prediction should be specified and a sample class should be described. The third type is class discovery, primarily by using unattended learning to identify a specific category based on their speech profiles' similarity.

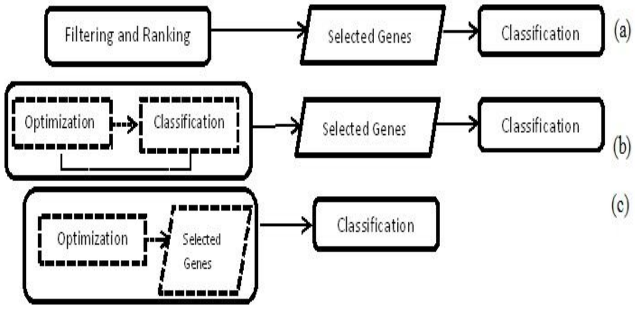

Figure 1: a) Filter Method b) Wrapper method c) Embedded Method [56]

The following section focuses on class prediction, genetic selection and methods of cancer classification.is often referred to as gene prioritization or biomarker discovery [9]. Due to the experiment phase , the data sparsity of the microarray exists. Many data sets with microarrays have missing post-computing values.

The research focuses on class predictions and gene selection procedures to identify the gene.There will be three types of characteristics checked, including filtering and wrapping methods using evolutionary biological methodologies.

2.3 Selection of characteristics (gene): Selection of characteristics involves finding genes expressed differently. This Filter, integrated, and wrapper methods [56] classifies the process of selecting features in three groups. The interplay with the classification model construction depends upon all of these methods. In the general context of feature choices new hybrid and ensemble approaches have been introduced recently. The next section and the algorithms of these three sections provide a summary.

(i)Filter approach: There is no clear learning pattern in the filter approach. Through this method different parameters were applied to the usual features (genes), while the features with the highest values were subsequently chosen. *Mutual information (MI),* measuring the dependency level between two random features, [11] is the most frequently used filter method. *Information gain (IG)* is a uniform filter approach measuring how often a feature provides information for a certain class[12]. *Minimum Redundancy Maximum Relevance (mRMR)* is a filter-based approach which selects features that maximize gene pertinence while the redundancies of each class[14]. *Symmetrical Uncertainty (SU)[*15] is built on information gains across a set of standardized values[0,1]. Selection of Ranks based on the association between experimental evaluations functions are based on (CfS) [15] Correlation-Based feature selection. CFS aims at reducing the amount of correlations between features and at the same time growing the feature in class correlations. The rapid filter for correlation (FCBF)[39] has been designed to distinguish both significant and redundant characteristics. This assesses individual features and recognizes dominant associations and heuristically eliminates redundant features. *Analysis Of Variance (ANOVA)* is also called Sum of Square BW[16].

The most commonly used measure of variance is Sum of Squares (SS). The Laplacian Score (LS)[18] is based on an unregulated filtering system, which can be related to the same class by two characteristics if nearby. Independent Component Analysis (ICA)[19] , a method of feature selection extracts independent variable using the linear representation of non-Gaussian results. Instance-based learning (IBL)[22] is a method of selecting features with a monitoring presence that undermines the need for a supervised recommendation. This is a method of controlled filtering that selects one feature based on a score of the fisheries criterion. This algorithm was also called as Fisher Score algorithm [17]. This seeks to optimize class differentiation and reduce shifts in classes to a minimum. The decision tree is a system of ranking based on the decision tree.

Ultimately, each tree has its own random vectors. The equivalent distribution applies to all the trees in the Random *ForestRanking (RFR)[20]*. By that the overall likelihood of top body,

(ii)Wrapper approach: the wrapper approach needs to be used to define the best subset of functions using learning methods[10]. For a closer result in the classification, the model hypothesis is combined into the search space. The wrapper technique's efficiency is determined based on the basic classifier precision. Typically, high calculation costs and an increased risk of over-fitting are required. The most commonly used wrapper methods are:

Heuristic search algorithm that encourages Natural development and process of natural selection [24]is the *Genetic Algorithm(GA)*. The selection, the crossover and the mutation of the algorithm have changed into three operations. Until passing to the next generation, the selection process picks the most appropriate chromosomes. Mutations are important to sustain a degree of population diversity. ABC is an evolutionary algorithm of Artificial bee colony[25] . It is a natural algorithm.

Ant Colony Optimization (ACO)[26] was initially motivated and developed by the ant colonies. The foundation of ACO is that ants will find from their colony the shortest path to the food source and vice versa. The ants then go to their colony randomly to look for food. The food quality is measured when ants locate a source of food. When this ant returns to the colony, it leaves pheromones on the way , leading other ants to the source of food. Thus, as the growing number of ants continues to realize the path is becoming stronger and stronger. The bird flocks, fish schooling patterns and theory of swarming are an inspiration for particle swarm optimization (PSO)[27] . Each PSO particle is a candidate solution that holds a certain location. Every particle bring up-to-date its location by adjusting its speed built on its history and its best success with its neighbor’s.

The Bat Algorithm (BA)[28]is a micro-bats-based natural algorithm, which utilizes the conduct of echolocation in locating its prow. At a certain speed and frequency, bats randomly migrate to different locations. A population-based optimization strategy is the Black Hole Algorithm (BHA)[29] . Black Holes behavior inspired by outer space. In the black hole everything near object disappears. BHA usually starts from the initial population of candidate solutions. Typical attention is given to the best solution for each iteration.

The Grasshopper Optimization Algorithm (GOA)[30] is nature-inspired and population-driven optimization algorithm. The behavior of grasshopper swarms is affected. Grasshoppers' positions are candidate solutions within this algorithm. A global optimization algorithm is used by *Firefly Algorithm*[31]. Through their blinking pattern, Fireflies attract other fireflies and their prey (the opposite sex). The algorithm was developed on the basis of three idealized rules: firstly, all fireflies, regardless of sex, are attracted to themselves.Third, by the type of objective feature the luminosity of the firefly is determined or affected. A meta-heuristic population-based algorithm inspired by the quest *Bhattacharya distance*[22] selects the most appropriate genes. The minimum error rate for the Bayes, the absolute maximum limit is expressed.

for genuinely harmonious state[32] is a harmonic-based search algorithm. Usually, musicians seek several different music pitch variations that are preserved in their minds while making harmonies. This HAS method comprises three stages. The harmonic memory is initialized, solutions improvised and the HM modified.

(iii)Hybrid/ Ensemble approach: The hybrid approach is for filters as well as wrappers. The approach of the ensemble assumes a better synthesis than one expert of many specialists' results. A single wrapper approach in one data set can easily produce fantastic results, and can also be terribly effective in another dataset. The hybridization of more than one method leads to a lower mean error rate. Nonetheless, the hybrid design can be less exact, as the filter and the wrapper are in separate stages [33] .

*Classification***:** It is a data mining method that allows the allocation (prediction) of the class label given to the collection of data. Classification is a type of supervised method which defines the class features. The following parts demonstrate the best classification methods in the wide field of Microarray data analysis. The precision of the classifier is then measured to also measure its precision.

SVM[34] supports the quest for a superfluous plane that divides the tuples optimally between groups. If the number of functions exceeds the number of samples, it works very well. In order to find hyper linear nonlinear surfaces, the process must expand from a linear SVM to a nonlinear SVM if the data is not linearly separate. This permits the use of other points to break the hyper plane, such that certain points can be misrepresented. By using high-dimensional data, this is efficient and fits well for vast volumes of data.

A simple instance-based, non-parametric, controlled learning algorithm, *K Nearest Neighbor KNN*[35] . The fundamental concept of K-NN requires the implementation of the metric similarity. In the K training data (most connected instances such as neighbors), new instances are found through a search procedure. The probabilistic classifier based on the theorem of Bayes is *Naive Bayes (NB)***[36]**. For each class, it measures the posterior likelihood with the likelihood of a class belonging to one particular attribute value. Bayesian belief networks were built that allows class dependency between variables to exist.

Genetic programming classification[41] is one of the common evolutionary algorithms (EA) and machine learning types. GP can be used in the dataset between genetic features, to find the functional relationship to a specific category. The common algorithm for classifying or grouping into a questioned subset is the Fuzzy Classification[43]. The member 's function and procuder is determined by a fake propositional procedure's real value. *Bagging Classifier*[48] is An algorithm that creates an original data set classifier for any random subset. Then they combine their respective predictions and classifications in order to make the final decision by voting or taking the average on the classification. *Neural networks (NN)**[51]*** are often referred to as neural artificial networks because they represent information treatment strategies in the human mind.

3. FEATURE SELECTION APPROACHES OF HYBRID METHOD

The selection of features is a central subject of study in data mining. The central goal of the selection of features is to eliminate noise and minor characteristics, and to choose the best and most reliable properties. This study attempts to use the selection method to choose the most insightful genes for the hybrid role. The hybrid approach focuses on two steps. Firstly, a preprocessing step for filtering noise off and secondly, wrapping techniques which use ideal functions to suit the subset. We will explore in this review a variety of approaches to the hybrid and ensemble using bio-inspired development methods. The success of such a system depends on the precise categorization and the selection of genes. We will review the current research on gene selection and cancer diagnosis using hybrid approaches.

3.1 Hybrid Approach: Most techniques for gene selection are not usage-by-filter approaches [13]. The risk of excess fitting and higher computational costs are due to this. Instead, the new work has widely used a combination of these approaches. In order to prevent high calculation costs and to make selection of the most insightful genes possible, it uses the filter method to pre-process the wrapper process.

Genetic algorithm (GA) - Hybrid Approach: *Authors in [37]*developed a new Hybrid algorithm of choice, known as *MIMAGA-* Selection, which combines *Mutual Information Maximization (MIM).* This was used to classify genes with high dependence on all other genes as a filter technique in the early MIM applications. Four different classification systems are also used to classify MIMAGA-Selection as the gene being selected: the neural back propagation network, the SVM (supporting vector machine), the ELM support system and the regularized extreme learning machine (RELM). More than 80% precision is achieved by all four classifications. The accuracy of MIM AGA selection is higher than the current function selection algorithms. The authors conclude that the algorithm is more accurate than current selection algorithms. The Microarray data analysis *in [38]* suggested a hybrid approach to the selection of features. The proposed approach uses a *Genetic algorithm for the Dynamic Parameter (GADP)* setting with the X2 check homogeneity. Next, a feature-selection method is used to pick 500 genes from the original sample as the BW ratio between groups and the sum of the square intergroup. To test the output of the proposed method against existing methods, six cancer data sets were used. The findings showed that GADP worked well beyond the current methods and in five datasets with fewer genes, 100 percent accuracy was achieved.

The new hybrid selection method, combining *Correlational-based Feature Selection (CFS) and Taguchi-Genetic Algorithm (TGA)***,** was proposed by *Chuang et al.[39]*. Two steps of the proposed process were introduced. To exclude unnecessary elements, the CFS filter method used to select correlation dependent characteristics. The next step was the TGA methodology for the functionality created from the filter stage and the best feature sub-set was established. The Taguchi genetic algorithm (TGA) has been used in the transformation and mutation cycle by the Taguchi-genetic algorithm (GA) and the Taguchi system. The K-nearest neighboring classifier (KNN) classification accuracy was applied in assessing the proposed method. The research was performed on 11 discrete cancer and multi-class datasets. The comparison revealed that the method proposed achieves the highest classification accuracy in 10 datasets where six datasets achieve 100 % accuracy.

In [40] authors suggested the Microarray data analysis focused on GA genetic algorithm and artificial intelligence offers a new evolutionary approach to genetic selection. The proposed approach was based on two phases. The first move was to pick filter methods for the 500 genes from Laplacian and Fisher. The next stage was the IDGA method and a random climb of the restart, which were focused on improving learning. Supporting devices have been employed in classification applications including SVM, Naïve Bayes (NBY) and KNN. The finding showed that on four datasets the proposed method had 100 % accuracy. For four datasets, Fisher's IDGA also overcomes Laplacian's IDGA. The feature selection system for classification using GE-SP, proposed by *Salem et al*.[41], was an *Integrated Information Gain (IG) and Standard Genetic Algorithm (SGA)* and evaluated in 7 Micro Array cancer datasets. Results showed that in two datasets, accuracy of 100% was achieved.

*Ant Colony Optimization - Hybrid Approach (ACO):*The hybrid methodology used in gene-selection to evaluate the micro array data is one of the methodologies suggested by *Sharbaf et al[42]i.e.Cellular Learning Automata with Ant Colony Optimization (CLA-ACO).* The plan comprised of 3 phases: The filter process was based on the fishing criterion system. The second stage was a hybrid mobile and anti-colony learning system. In the third step, the final subset of features from the second phase were classified in the subset. The evaluation of the proposed method was carried out in two phases. The first step is to evaluate the different classification methods for selecting the features. Four classification methods were used: T-testing, collecting knowledge and Fisher and Z-score. The best classification method was the Fisher test. In the second part, the CLA-ACO model is evaluated. Four binary and multi-class cancer datasets were used. The proposed CLACOFS has been performed 20 times and the average accuracy has been calculated. Three equivalent classifications were the SVM, KNN & Naïve Bayes. The outcomes indicate that the Naïve Bayes classification is unlike other classifications. This approach also achieved 100 percent precision in two data sets. In addition, in two datasets, it was suggested*[63]*.

The contribution to the creation of a Microarray Analysis Data Classification Scheme for *hybrid stem cell (HSC)* is *Vijay and Ganesh Kumar[43]*. Optimizing the ant colony (ACO) & novel Adaptive Stem Cell Optimization (ASCO) were used in the proposed method. The Mutual Information (MI) technique was used to select insightful genes during pre-processing. Five microarray data sets were tested to assess the efficiency of the proposed system. The precision was opposed to the other, for example, *Hybrid Colony Algorithm (HCA)* fuzzy-based classification systems. Both techniques were applied by the proposed system. The results showed.

Bat Algorithm (BA) - Hybrid Approach: In *[44]* they proposed this algorithm; the BAT algorithm was multifunctional with sophisticated specification, MOBBA-LS and the proposed algorithm, multi-objectiveness operators and effective local search strategies. Fisher criteria were used to pick the upper 500 genes in the original filter technique. This was then used in the MOBBA-LS process as the filtered subset. Three Microarray cancer data sets were included. Significantly fewer genes than the comparable state of the art methods for the processing of Prostate information achieved the highest recorded accuracy. The four classifiers: Leave One-Out Cross validation (LOOCV) and Decision Tree (DT) were used to test any subset of validated support- vector machine (SVM), K-Nearest Neighbors, Naive Bayes (NBY) and Decision-Tree.

*Artificial bee colony - Hybrid Approach***:** In *[45]* proposed a new technique for gene selection for microarray studies for hybrid feature selection. Informative genes were identified using Independent Component Analysis (ICA) as well as Artificial Bee Colony (ABC). The ICA picked an average of 50 to 180 genes from initial data sets. Initially the tests showed that the maximum classification precision in ICACABC algorithms is reached in four datasets in comparison with the other methods of gene selection existing. The algorithm proposed was used 30 times. ICACABC analyzed six microarray benchmarks for cancer*[61]*. For estimate the accuracy of classification in Naïve Bayes, Leave One Out Cross Validation (LOOCV) was used.

Authors in *[46]* . A new gene selection algorithm, together with an ABC artificial colony algorithm known as mRMR-ABC, is being proposed. Finally, the Support Vector Machine (SVM) classification was used to calculate the effectiveness of the proposed solution[67]. AmRMR-ABC study is conducted using six cancer datasets with binary and multi-class gene expression. The proposed algorithm (mRMR-GA) and (mRMR-PSO) have been compared in accordance with the existing gene selection methods. For five cancer datasets, the algorithm achieved 100 % accuracy*[59]*.

*Particle Swarm Optimization (PSO) - Hybrid Approach***:** In *[23]*theyproposed a new hybrid search algorithm, based upon HPSO-LS and Particulate Swarm Optimization (PSO). In selecting specific features, the objective was to use correlation data features for guiding the pso search method. The efficiency of the proposed approach was assessed by 12 cancer datasets. The experiments included the neighbor (k-NN)[66]. The exactness of the technique proposed was contrasted with previous wrapper-based performance assessment approaches.Microarray data suggests hybrid gene therapy by *Jain et al.[47],* the combined correlation-based feature selection approach (CFS) was used to compare the result with the seven existing selection processes, in combination with the improved binary particle-swarm optimization (iBPSO*)*. In 10 datasets, the system achieved the highest accuracy and 100% accuracy in 7 datasets. For test the proposed procedure, the 10-fold cross-validation classification Naïve Bayes has been used. The highest precision in 10 data sets and 100 percent precision in seven data sets has been achieved by CFS-iBPSO*[62]*.

Black Hole Algorithm – Hybrid Approach: *Authors in.[48]* , the new gene selection technique was proposed as a filter method for the microarray results based on RFR data from the Binary Black Hole Algorithm (BBHA) and Random Forest Ranking (RFR).Four Microarray cancer data were used to test the proposed approach in the 10-fold cross validation Bagging classifier[69]. The proposed approach was 100 times used. The method was compared with seven known classifiers and found to be most accurate.

Biogeography Algorithm- Hybrid Approach: *Li and Yin[49]* suggested the multi-objective binary biogeography (MOBBO) approach to gene selection. Biogeography As a first step in selecting the top 60 genes, the selector Fisher-Markov was applied. Ten Microarray datasets have been tested and three PSO-based methods have compared the result. In nine data sets, three of which were 100 percent obtained, MOBBOCSVM achieved maximum accuracy. In *[50]* proposed hybrid gene-selection filter wrapper methods in the Microarray Data *Harmony Search Algorithm (HSA)*. Combined approach with a Harmony Search Algorithm (HSA), Symmetric uncertainty (SU) known as SU-HSA. Two classifiers, IB1 and Naive Bayes (NB), were used to test the proposed system. For 10 microarray datasets, the experiment was replicated 10 times. The new approach is most accurate in 5 data sets and 100% accurate in four of them compared with state-of-the-art gene selection*[60]*.

GOA (Grasshopper optimization) - Hybrid Approach: *Tumuluru and Ravi [51]* proposed Deep Belief neural networks (DBN-based) based on Grasshopper Optimization to pick features in microarray tests. The results are based on Grasshopper optimization. In the first phase of pre-treatment, data processing involved two phases and in the second phase, gene selection. The distance from the gene was used to select the appropriate characteristics and to delete redundant characteristics during gene selection*[57]*. In comparison to three existing approaches, two data sets from Microarray were checked. The proposed process reaches the greatest exactness in accordance with the alternatives.

3.2 Ensemble Approach: *[13]*It is based on the premise that the performance of many experts is greater than the performance of a single expert. This method has recently been applied to problems related to Microarray genes collection and data classification[65]. The following section addresses different algorithms using the ensemble approach with naturally-inspired evolutionary methods. A meta heuristic system for genetic selection with *Harmony Search(HS)* has been proposed by *authors in[52]*. The model proposed was based on two phases.

During the first step, the HS method was included in the GA cycle and the random solutions were graded according to their fitness value. In phase two (new solution), GA picked the top-ranking solution to generate the offspring. Seven methods of selecting features were used to analyze the current model. There was also a lower error rate with greater accuracy*[58]*.

Authors in [53] suggested the Microarray data collection of genes to provide a *Binary Particle Swarm Optimization (BPSO) Hybrid Approach &CGA*. The compact GA is built into BPSO for every generation and acts as a local optimizer[64]. The compact GA was used to replicate, crossover and mutate the particles produced. As a group, the neighbor of K-nearest (K-NN).Ten Microarray data sets were tested for the proposed method. The result showed that in nine datasets, the proposed approach achieved the lowest error rate. Two hybrid approaches were introduced and compared by *Djellali et al.[54].FCBF Filter with GM algorithm* (FCBF -GA) Fast Correlations were the first method; *FCBF with PSO Particle Swarm Optimization (FCBF-POS)* FCBF was the second method. The precision and number of genes selected for FCBF-POS were higher than the FCBF-GA ranking. Four microarray data sets and the SVM support system were used to evaluate the proposed methods. The genetic algorithm (GA) was combined with the Artificial Bee Colony algorithm (ABC).
Al-shamlan and other[55] Proposed Genetic Bee Colony (GBC) new hybrid gene selection approach. Six Microarray datasets, both binary & multi-class, made the performance evaluation of the proposed algorithm. The precision of classification was calculated with the SVM classifier / for each data set the evaluation experiment was performed 30 times. In addition to the recently published gene selection procedures, the algorithm (mRMR-ABC), (mRMR-GA) and (mRMR-PSO) were compared[68].

Table 1: shows the efficiency of the different hybrid selection approaches for particular data sets for gene expression in current literature.

Original Reference	Hybrid Methodology	Classifier	Datasets Used	Classification Accuracy in percentage	Number of selected genes
[43]	Hybrid Stem Cell (HSC) (Mutual Information and ACO)	Fuzzy Classification	Colon, Leukemia1 Prostate	100 ,100 90.85	-- --
[44]	fisher criterion, Bat algorithm (BA)	SVM	SRBCT,Prostate	85 , 94.1	6 , 6
		KNN		100 , 97.1	6 ,6
		NB		100, 97.1	6, 6
[47]	CFS-iBPSO	NB	Colon, SRBCT Leukemia1,Leukemia2 Lymphoma, MILL Breast	94.89. 100, 100,100 ,100, 100, 100	4, 34, 4 6, 24 ,30 10
[54]	FCBF-POS and GA	SVM	Colon , DLBCL	96.30 100	1000 3204
[40]	Laplacian and Fisher score, Genetic Algorithm	SVM	SRBCT Leukemia 1 Prostate Breast DLBCL	100, 100 , 96.3, 100 ,100	18, 15, 14, 2, 9
		KNN		91.6, 97.2, 95.6,95.5, 97.9	-- --
		NB		98.2, 93.1 93.4,100, 95.8	-- --
[41]	IG and SGA	GP	Colon, Leukemia 1 Lung, Prostate	85.48, 97.06, 100, 100	60, 3 ,9 26
[51]	Logarithmic transformation, Grasshopper Optimization Algorithm	NN	Colon Leukemia 1	95 94	-- --
[42]	CLA-ACO and fisher criterion	SVM	Leukemia 1 Prostate	95.95, 98.35	3 , 14
		KNN		94.30, 99.25	3 , 15
		NB		95.95 , 99.40	4 , 10
[45]	ICA and ABC	NB	Colon ,Leukemia 1 Leukemia 2 , Lung	98.14 ,98.68 97.33 ,92.45	16 ,12 15 ,24
[46]	mRMR-ABC	SVM	Colon, SRBCT Leukemia1,Leukemia, Lung, Lymphoma	96.77, 100 100 ,100 100 ,100	15, 10 14 ,20 8 ,5

[39]	CFS and TGA	KNN	SRBCT , Prostate Lung	100 , 99.22 98.42	29 ,24 195
[23]	Probabilistic random function, Particle Swarm Optimization (PSO)	KNN	Colon, Leukemia 1 Lymphoma	84.38 ,89.28 87.71	60 ,100, 50
[48]	RFRBBHA	Bagging Classifier	Colon, MILL	91.93, 98.61	3, 5
[49]	Fisher-Markov selector, Multi-Objective Binary Biogeography (MOBBBO)	SVM	SRBCT, Prostate Lung	100, 98.3, 98.4	6, 12 ,16
[50]	SU with HSA	NB	Colon, SRBCT, Leukemia1,Leukemia2 Lymphoma ,MILL	87.53,99.89,100 ,100,100 ,98.97	9 ,37 26 ,24 10 ,10
[37]	Mutual Information Maximization Adaptive Genetic Algorithm (MIMAGA)	SVM	Colon	83.41	202
[38]	GADP	SVM	Colon, DLBCL SRBCT, Leukemia1	100, 100 100 ,100	8, 6 8 ,5
[55]	mRMR, GBC and GA	SVM	Colon, SRBCT Leukemia1,Leukemia2 , Lung,Lymphoma	98.38, 100 100 ,100 100, 100	10, 6 4 ,8 4 ,4

4.CONCLUSION

The study of microarray data provides useful results to solve problems of gene expression. Cancer detection is oneof the key applications in the Micro Array data studies. The classification is difficult because of the high dimensionof a limited gene expression sample. Therefore, a feature selection technique is the most practical method forovercoming these challenges. The selection of features is important for classification in large volumes of data. Theunique nature of microarray data (large genes but few specimens) makes gene selection an important need. Severalexperiments in feature selection are under way to diagnose cancer with gene expression data from microarrays.

A variety of hybrid algorithms using a bio-inspired wrapper technology were used in Microarray data analysis forgene selection and cancer classification. We have conducted this analysis to analyze and compare these algorithms.The genetic algorithm is the best commonly used wrapper technique in the literature. Genetic algorithm amongother methods is applied with the maximum precision with comparatively few selected genes.

REFERENCES

[1] PM Narendra, K Fukunaga**A branch and bound algorithm for feature subset selection**. *IEEE Trans. Comput.*, vol. C-26, no. 9, pp. 917-922. 1977.

[2] H Alshamlan, G Badr, et al **A comparative study of cancer classification methods using microarray gene expressionprofile**. in*Proc.1st Int. Conf. Adv. Data Inf. Eng. (DaEng). Singapore: Springer* pp. 389-398. 2014.

[3] MM Babu, **Introduction to microarray data analysis. In Computational Genomics:Theory and Application** pp.225-249. 2004

[4] J Read,S**Brenner Microarray technology.***In Encyclopedia of Genetics*. New York, NY, USA: Academic, p. 1191. 2001

[5] A. Perez-Diez, A. Morgun, and N. Shulzhenko, **Microarrays for Cancer Diagnosis and Classication**.*Austin, TX, USA:Landes Bioscience*, 2013.

[6] R Simon, **Analysis of DNA microarray expression data**. *Best Practice Res., Clin. Haematol.*, vol. 22, no. 2, pp.271-282. 2009.

[7] C Gunavathi, K Premalatha, et al **A survey on feature selection methods in microarray gene expression data for cancer classification**.*Res.J.PharmacyTechnol.*,vol.10,no.5,pp.1395-1401.2017.

[8] AL Tarca, R Romero, et al **Analysis of microarray experiments of gene expression profiling.***Amer. J. ObstetricsGynecol*., vol. 195, no. 2, pp. 373-388, 2006.

[9] C Lazar, J Taminau, et al, **A survey on filter techniques for feature selection in gene expression microarray analysis**. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 9, no. 4, pp. 1106-1119, 2012.

[10] Y Saeys, I Inza, et al. **A review of feature selection techniques in bioinformatics**.*Bio info*,vol.23,no.19,pp.2507-2517, 2007.

[11] JR Vergara, PA Estévez, **A review of feature selection methods based on mutual information**.*Neural Computing*.Appl., vol. 24, no. 1, pp. 175-186, 2014.

[12] ZM Hira , DF Gillies**A review of feature selection and feature extraction methods applied on microarray data**. *Adv.Bioinformatics*, vol. Art.no. 198363.

[13] V Bolón-Canedo, N Sánchez-Maroño, et al, **A review of microarray datasets and applied feature selection methods.***Inf. Sci.*, vol. 282, pp. 111-135, 2014.

[14] H Peng, F Long,et al, **Feature selection based on mutual information criteria of max-dependency, max-relevance, andmin-redundancy**. *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 12261238, 2005.

[15] MA Hall, **Correlation-based feature selection for machine learning**.*Univ.Waikato,Hamilton,New Zealand, Tech. Rep.* 1999.

[16] LS Kao , CE Green **Analysis of variance: Is there a difference in means and what does it mean?***J. Surgical Res.*, vol.144, no. 1,pp. 158-170, 2008.

[17] Q Gu, Z Li, et al, **Generalized Fisher score for feature selection**. *arXiv*:1202.3725, 2012.

[18] X He, D Cai, et al, **Laplacian score for feature selection**. *in Proc. 18th Int. Conf. Neural Inf. Process. Syst.*, p. 8, 2016.

[19] A Hyvärinen, EA Oja, **fast xed-point algorithm for independent component analysis.***Neural Comput.*, vol.9,no. 7,pp. 1483-1492, 1997.

[20] L Breiman ,**Random forests**. *Mach. Learn.*, vol. 45, no. 1, pp. 5-32, 2001.

[21] DW Aha, D Kibler, et al, **Instance-based learning algorithms**. *Mach. Learn*., vol. 6, no. 1, pp. 37-66, 1991.

[22] G Xuan, X Zhu, et al, **Feature selection based on the Bhattacharyya distance**. *In Proc. 18th Int. Conf. PatternRecognition.*, vol. 3, Washington, DC, USA, p. 957, 2006.

[23] P Moradi , M Gholampour, **A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy.***Appl. Soft. Comput.*, vol. 43, pp. 117-130, 2016.

[24] J McCall, **Genetic algorithms for modeling and optimization**. *J. Com-put. Appl. Math.*, vol. 184, no. 1, pp. 205-222, 2005.

[25] D KarabogaAn idea based on honey bee swarm for numerical optimization. Dept. Comput. Eng., ErciyesUniv.,Kayseri, Turkey,Tech. Rep. TR06, p. 10, 2005

[26] M Dorigo, V Maniezzo, et al, **Ant system: Optimization by a colony of cooperating agents**. *IEEE Trans. Syst., Man,Cybern. B, Cybern.*, vol. 26, no. 1, pp. 29-41, 1996.

[27] J Kennedy , R Eberhart, **Particle swarm optimization**. *In Proc. IEEE Int.Conf.Neural Networks.*,vol. 4,pp. 1942-1948. 1995.

[28] XS Yang ,**A new meta heuristic bat-inspired algorithm. In Nature Inspired Cooperative Strategies for Optimization***(NICSO 2010). Berlin, Germany: Springer*, pp. 65-74,2010.

[29] A Hatamlou, **Black hole: A new heuristic optimization approach for data clustering**. *Inf. Sci.*, vol. 222, pp. 175184, 2013.

[30] S Saremi, S Mirjalili, et al, **Grasshopper optimization algorithm: Theory and application**. *Adv. Eng. Software.*, vol.105, pp. 30-47,2017.

[31] XS Yang ,**Firefly algorithms for multimodal optimization**. *In Proc. Int. Symp.Stochastic Algorithms. Berlin, Germany: Springer*, pp. 169-178, 2009.

[32] ZW Geem, JH Kim, GV Loganathan, **A new heuristic optimization algorithm:Harmony search**.*J.Simul.*,vol.76,no. 2,pp. 60-68, 2001.

[33] A Jovi¢, K Brki¢, et al, **A review of feature selection methods with applications**. *In Proc. 38th Int. Conv. Inf.Commun.Technol., Electron.Microelectron.(MIPRO)*, pp. 1200-1205, 2015.

[34] J Han, J Pei, et al, **Data Mining: Concepts and Techniques.***Amsterdam, The Netherlands: Elsevier*, 2011.

[35] T Cover , P Hart, **Nearest neighbor pattern classification.***IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21-27,1967.

[36] S Taheri , M Mammadov**, Learning the naive Bayes classifier with optimization models**. *Int. J. Appl. Math. Comput.Sci.*,vol. 23, no. 4,pp. 787-795, 2013.

[37] H Lu, J Chen, et al ,**A hybrid feature selection algorithm for gene expression data classification**. *Neuro computing*, vol.256, pp. 56-62,2017.

[38] CP Lee , Y Leu , **A novel hybrid feature selection method for microarray data analysis**. *Appl. Soft Comput*., vol. 11, no. 1, pp. 208213. 2011.

[39] LY Chuang, CH Yang, et al, A hybrid feature selection method for DNAmicroarraydata.Computer.Biol.Med,vol.41,no.4,pp.228-237, 2011.

[40] M Dashtban, 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.

[41] H Salem, G Attiya, et al., **Classification of human cancer diseases by gene expression profiles**. *Appl. Soft Computing*., vol. 50,pp. 124-134,2017.

[42] FV Sharbaf, S Mosafer, et al, **A hybrid gene selection approach for microarray data classification using cellularlearning automata and ant colony optimization**; *Genomics*, vol. 107, no. 6, pp. 231-238,2016 .

[43] SAA Vijay, PG Kumar, **Fuzzy expert system based on a novel hybrid stem cell (HSC) algorithm for classification ofmicro array data**. *J. Med. Syst.*, vol. 42, no. 4, p. 61. 2018.

[44] M Dashtban, M Balafar, et al, **Gene selection for tumor classification using a novel bio-inspired multi-objective approach.***Genomics*, vol. 110, no. 1, pp. 10-17,2018.

[45] R Aziz, CK Verma, et al,**A novel approach for dimension reduction of microarray**.*Comput.Biol.Chem.*,vol.71, pp.161169,2017.

[46] H Alshamlan, G Badr et al, **mRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling.***Biomed. Res.Int.*,vol.2015, Art. no. 604910, 2015.

[47] I Jain, VK Jain, et al, **Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification.***Appl. Soft Computing.*, vol. 62, pp. 203215,2018.

[48]Pashaei , M Ozen, et al, **Gene selection and classification approach for microarray data based on random forest ranking and BBHA.***In Proc. IEEE-EMBS Int. Conf. Biomed. Health Inform. (BHI)*, pp. 308-311,2016.

[49] X Li , M Yin, Multi objective binary biogeography based optimization for feature selection using gene expression data. IEEE Trans. Nano bioscience.,vol. 12, no. 4, pp. 343-353. 2013. [50] SS Shreem, SS Abdullah x, et al .**Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm.***Int. J. Syst. Sci.*, vol. 47, no. 6, pp. 1312-1329, 2016.

[51] P Tumuluru, B Ravi, **GOA-based DBN: Grasshopper optimization algorithm-based deep belief neural networks for cancer classification.***Int. J. Appl. Eng. Res*., vol. 12, no. 24, pp. 14218-14231. 2017.

[52] K Das, D Mishra, et al, **A Meta heuristic optimization framework for informative gene selection.***Inform. Med.Unlocked*, vol. 4, pp. 1020, 2016.

[53] LY Chuang, CH Yang, et al, **A hybrid BPSOCGA approach for gene selection and classification of microarray data**. *J.Comput. Biol.*, vol. 19, no. 1, pp. 68-82,2011.

[54] H Djellali, S Guessoum, et al ,**Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection**. *In Proc. 5th Int. Conf. Elect. Eng.-Boumerdes (ICEE-B)*, pp. 1-6, 2017.

[55] HM Alshamlan, GH Badr, et al , Genetic bee colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Comput. Biol. Chem., vol. 56, pp. 49-60. 2015.

[56] P Yang, BB Zhou, et al, **A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data**. *BMC Bioinf*., vol. 11, no. 1, p. S5,2010.

[57]. P Amulya, S SaiMeghana , A Manisha , P Rajarajeswari, **A Deep Learning Approach For Brain Tumor Segmentation using Convolution Neural Network**, *International Journal of Scientific and Technology Research,*December 2019.

[58]. R Rita Kamble, Dr.P. Raja Rajeswari, **A review of various camouflage moving object detection techniques,***Journal of Engineering and Applied Sciences*, 2017.

[59]. Navya Krishna , P Rajarajeswari et al, **Recognition of Fake Currency Note using Convolutional Neural Networks** ,*International Journal of Innovative Technology and Exploring Engineering (IJITEE)* ISSN: 2278-3075,Volume-8 Issue-5 March, 2019.

[60]. T.Ganesan, PothurajuRajaRajeswari, **Genetic Algorithm Based Optimization to Improve the Cluster Lifetime by Optimal Sensor Placement in WSN's,***International Journal of Innovative Technology and Exploring Engineering (IJITEE)* , ISSN: 2278-3075, Volume-8 Issue-8,June, 2019.

[61]. A.Santosh. Shinde , Dr. P. Raja Rajeswari ,**A Novel Hybrid Framework for Cuff-Less Blood Pressure Estimation based On Vital Bio Signals processing using Machine Learning**, *International*

Journal of Advanced Trends in Computer Science and Engineering, Volume 9 No.2, March -ApriL 2020.

[62]. K.SaiSravani,Dr.P.RajaRajeswari, **Prediction Of Stock Market Exchange Using LSTM Algorithm**, I*nternational journal of scientific technology research*, VOLUME 9, ISSUE 03, MARCH 2020.

[63]. YanishPradhananga, PothurajuRajarajeswari, **Tiarrah Computing: The Next Generation of Computing**, *International journal of Electrical and Computer Engineering* 81(2),pp. 1247- 1255. 2018.

[64] Dr.P.RajaRajeswari,P.SuryaTeja,P.SriHarsha,T.Chaitanya Kumar, **Enhancing the Performance of Crime Prediction Technique Using Data Mining,***International Journal ofEngineering & Technology*, 7 (2.32) 424-426,2018.

[65]Dr.P.RajaRajeswari, Supriyamenon.M, **A contemporary way for enhanced modeling of context aware privacy system in PPDM**. *Journal of Advanced Research in Dynamic and Control systems*,Vol.10,01-issue,July 2018.

[66] M.Supriyamenon , Dr. P.Rajarajeswari,**A Review on Association Rule Mining Techniques with Respect to their Privacy Preserving Capabilities**,*International Journal of Applied Engineering Research*, Volume 12, Number 24 pp. 15484- 15488. july 2017.

[67]T.Ganesan, PothurajuRajaRajeswari,**Genetic Algorithm Approach improved by 2D lifting scheme for sensor node placement in optimal position** , *IEEE xplore*, ISBN: 978-1-5386-7798-8978. November 2019 .

[68] Ahmad al-Qerem,ArwaAlahmad, **Human Body Poses Recognition Using Neural Networks with Data Augmentation,***International Journal of Advanced Trends in Computer Science and Engineering*,vol.8, No.5, September-October 2019.

[69] VyacheslavLyashenko, Syed Khalid Mustafa, SvitlanaSotnik, M. Ayaz Ahmad,**Basic Principles of Decision Making upon Receipt of New Nanomaterial,** *International Journal of Advanced Trends in Computer Science and Engineering*, vol.8, No.5, September-October 2019.