IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 12, 2019, accepted June 3, 2019, date of publication June 14, 2019, date of current version June 28, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2922987

A Survey on Hybrid Feature Selection Methods in Microarray Gene Expression Data for Cancer Classification

NADA ALMUGREN AND HALA ALSHAMLAN
Information Technology Department, King Saud University, Riyadh 11362, Saudi Arabia
Mechanical Engineering Department, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

Corresponding author: Hala Alshamlan (halaa@mit.edu)

This work was supported by a grant from the Research Center of the Center for Female Scientific and Medical Colleges, Deanship of Scientific Research, King Saud University.

ABSTRACT The emergence of DNA Microarray technology has enabled researchers to analyze the expression level of thousands of genes simultaneously. The Microarray data analysis is the process of finding the most informative genes as well as remove redundant and irrelevant genes. One of the most important applications of the Microarray data analysis is cancer classification. However, the curse of dimensionality and the curse of sparsity make classifying gene expression profiles a challenging task. One of the most effective methods to overcome these challenges is feature (gene) selection. In this paper, we aim to review and compare the most recent hybrid approaches that employ bio-inspired evolutionary methods as the wrapper method.

INDEX TERMS Microarray, gene selection, bio-inspired, hybrid approach, cancer classification, gene expression.

I. INTRODUCTION

DNA Microarray technology (also known as DNA chips or gene chips) is a powerful tool that helps researchers monitor the gene expression level in an organism. Microarray data analysis provides valuable results which contribute towards solving gene expression profile problems. One the most important applications of Microarray data analysis is cancer classification. Cancer may be a genetic disease; the analysis of cancer pathobiology is the analysis of genes that cause cancer, i.e. the gene whose mutation is responsible for cancer. This reflects the changes in the expression level of various genes. However, classifying the gene expression profile is a challenging task and considered as (NP)-Hard problem [1]. Hence, not all genes contribute to the presence of cancer. A vast number of genes are irrelevant or insignificant to clinical diagnosis. Therefore, incorrect diagnoses can be reached when all the genes are used in Microarray gene expression classification. There are two main issues related to the analysis of Microarray data; first, the dataset in Microarray is high-dimensional which means it contains several thousand genes (features) and it has low data sparsity, meaning it has a low number of samples, usually tens of samples. Second, gene expression data has a high complexity; genes are directly or indirectly correlated to each other. Standard machine learning methods did not perform well because these methods are best suited when there are more samples than features.

In an attempt to overcome these issues, dimension reduction or feature (gene) selection algorithms have been applied. Generally, the gene selection methods are categorized into three categories: filter, wrapper, and embedded methods. The filter method involves each feature being evaluated individually by using its general statistical properties. The wrapper method uses learning techniques to select the optimal feature subset. The quality of the wrapper technique is estimated by the accuracy of the specific classifier. The wrapper approach usually employs evolutionary or bio-inspired algorithms to guide the search process. The embedded method searches for the optimal feature subset and is built in the classifier; the search space is combined in the hypothesis space. Recently, hybrid and ensemble methods were added to the general framework of feature selection. A hybrid approach is built to take advantage of both filter and wrapper approaches. Thus, it combines the computational efficiency of the filter approach with the high performance of the wrapper approach.

The associate editor coordinating the review of this manuscript and approving it for publication was Yucong Duan.

Bio-inspired evolutionary methods have widely applied the wrapper approach in feature selection for Microarray data analysis and demonstrate a superior performance [2]. However, this study is aiming to review and compare the most recent hybrid approaches that employ bio-inspired evolutionary methods as the wrapper method. Two steps are central to the application of the hybrid approach. The first involves a pre-processing step applied to filter off noise and the second involves wrapper techniques that homes in on the subset employing optimum features. The performance of such an approach is dependent on two factors; the classification accuracy and the number of selected genes.

II. BACKGROUND

In this section we will review basics concepts of Microarray technology and Microarray data analysis. First, we will give an overview of Microarray technology and the Microarray gene expression profile. Then, we will discuss Microarray data analysis and its types. This study will focus on class prediction (classification), as well as its methods for gene selection cancer classification. Three categories of feature selection will be reviewed including filter methods and wrapper methods that employ bio-inspired evolutionary methods.

A. MICROARRAY GENE EXPRESSION PROFILE

DNA Microarray technology (also called 'DNA chips') has become a powerful tool for biologists to monitor the gene expression levels within an organism [3]. This technology enables researchers to simultaneously measure the expression levels of a large number of genes. Gene expression data generally includes thousands of genes (high dimensionality), as well as a small number of samples. It also contains numerous irrelevant and redundant features. Microarray technology is used most within medical fields, in order to learn about what causes diseases and how to treat them [4]. Researchers have figured out that the mutations in DNA may sometimes be the cause for certain diseases, like breast cancer. The mutation of certain known genes, is known as being the cause for some diseases. However, there is no one type of mutation that causes all diseases [79]. DNA Microarray data analysis is therefore used in order to discover and detect general mutations within DNA.

Microarray gene expression data technology has had a massive effect on cancer research. It is a powerful technique when it comes to diagnosing and identifying the disease genes for human cancers [5]. Moreover, it has been vastly used to identified cancer-related genes using feature selection methods [6].

B. MICROARRAY DATA ANALYSIS

Gene expression data analysis is the process of finding the most informative genes, whilst also removing any redundant and irrelevant genes [7]. Three types of Microarray data analysis currently exist, which are: class comparison, class prediction, and class discovery [8]. Class comparison is also known as gene discovery selection, as it typically focuses on defining the genes that are differentially expressed between predefined classes [8]. Class prediction is also known as classification, or supervised learning, and it involves finding the class of a sample. The third type is class discovery, or clustering, which is basically when unsupervised learning is used to find a related group based on the similarity of their expression profiles. The next section will focus on class prediction, as well as its methods for gene selection and cancer classification.

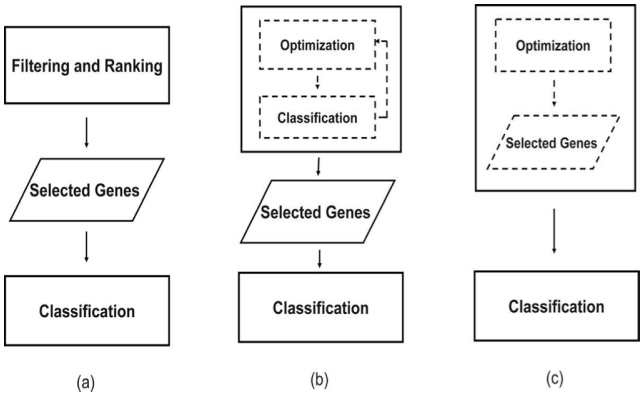

FIGURE 1. The different between feature selection methods a) filter, b) wrapper, c) embedded. [80].

1) FEATURE (GENE) SELECTION

In gene expression Microarray data analysis, feature selection techniques are typically used to find the informative genes. Feature selection is how differentially expressed genes are discovered [9]. The process of feature selection is also called gene prioritisation, or biomarker discovery [9]. Microarray data analysis process is challenging task, as there are ultimately too few samples, that in turn have too many features. The data sparsity of microarray exists due to the process of experiments. Many microarray data sets have missing values which affects the post-processing. Some methods are used to deal with it, such as a Singular Value Decomposition (SVD) based method (SVDimpute), weighted K-nearest neighbors (KNNimpute), and row average [78]. Feature selection methods are categorised into three categories, which are: filter, embedded, and wrapper methods [80]. These methods all depend on how they will interact with the construction of the classification model [10]. Recently, new hybrid and ensemble methods have been added to the feature selections general framework. In figure 1, we present a logical diagram to show the relationship between filter, embedded, and wrapper approaches. The next section offers an overview of these three categories, along with their algorithms.

a: FILTER APPROACH

In the filter approach, each feature is evaluated individually via using its general statistical properties [9]. The filter approach does not use any specific learning model. Thus, it is independent of the classifier. In this approach, the typical features (genes) were ranked via specific criteria, while the

features with the highest scores were then selected. Following this, these features then get used as input for the classifier or wrapper methods. The most typically used filter methods are:

Mutual Information (MI) which measures the level of dependence between two random features. In other words, this process measures the amount of information that one variable (X) knows about another variable (Y) [11].

Information Gain (IG) [12], is a uni-variant filter approach, which measures how much information pertaining to a certain class a feature offers. Therefore, the feature offering the most information is highly related, while the unrelated feature offer no information. Information gains are measured in the base of entropy (level of impurity). A threshold is thus set, with any feature that appears higher than the threshold then being selected. The higher the information gain, the higher class purity, which ultimately results in a high chance of getting the target class [13].

Minimum Redundancy Maximum Relevance (mRMR) [14], is a multi-variant filter method, which involves selecting the features that maximise the genes' relevance, while reducing the redundancies within each class. Thus, any mutually exclusive features not mimicking each other are selected.

Symmetrical Uncertainty (SU) [15], is a simple procedure that is used to evaluate the goodness of the features. This is basically a normalised version of Mutual Information (MI). SU is based on information gains, via the normalised values from within the range [0,1]. Symmetrical essentially means that SU treats the feature symmetrically, with an example being that SU (X, Y) is equal to SU (Y, X). This ultimately means that the number of comparisons is reduced.

Correlation-Based

Feature Selection (CFS) [15], is a multi-variant filter method that ranks features, based on the correlation between the heuristic evaluation functions. The aim of CFS is to reduce the amount of feature to feature correlations, while increasing the feature to class correlations.

Fast correlation-based filter (FCBF) [50], is a multivariate gene selection method that starts with a full set of features (genes). It uses the symmetrical uncertainty (SU) measurement to calculate dependencies of genes and _finds the best subset using a backward selection technique with a sequential search strategy. FCBF is a computationally efficient algorithm designed to identify both irrelevant and redundant features. It evaluates individual attributes and identifies predominant correlations, then heuristically removes redundant features. It has an inside stopping criterion that makes it stop when there are no features

Between-Groups

to Within-Groups Sum of Square BW [16], is also called Analysis Of Variance (ANOVA). ANOVA aims to identify any potential significant differences between the means of two or more groups (classes). Therefore, it measures between the group variations and within the group. Sum of Squares (SS) is the variation index that is most used. The features with the highest BW are selected.

Scoring algorithm [17], is also known as the Fisher Score algorithm, which is a supervised filter technique method that selects a feature based on the fisher criterion score. The chief purpose of the Fisher score is identifying feature subsets where the distances between the data points within different classes are large, while there are small data points within the same class. One could also say it seeks to maximise the separation between the classes, whilst minimising any variations within the classes [18].

Laplacian Score (LS) [19], is based on unsupervised feature filter method, with two features possibly being related to the same class should they be close to each other. LS is then based on the Laplacian Eigenmap and Preserving Projection. Thus, every feature is then judge according to its power to preserve its own locality.

Independent Component Analysis (ICA) [20], [21], is a feature selection method that finds the linear representation of non-Gaussian data, so that the independent component is extracted. It therefore decomposes the features when every feature is statistically independent of one another.

Random Forest Ranking (RFR) [22], is a method for feature ranking that is based on the decision tree. Each tree is essentially dependent on independent random vectors. In RFR, all of the trees have the same distribution.

Instance-Based

Learning (IBL) [23], is a feature selection characterised by the presence of supervision, which depends on the instance-based feature ranking. Each instance contains a certain number of features. In each instance, the features are ranked according to their weight. Different instances then provide different feature rankings. The top-ranked features within each instance are ultimately selected.

Bhattacharya distance [24], selects the highly relevant genes, by minimising the overall probability of upper bound. The minimum upper bound for the error is expressed in terms of a minimum Bayes error rate. The Bhattacharyya distance that is calculated between the classes uses co-variance as a factor, along with the vector mean of each class [25].

b: WRAPPER APPROACH

The wrapper approach concept involves using learning techniques to select the optimal feature subset [10]. The model hypothesis combines with the classifier in the search space, in order to reach a more accurate classification result. The wrapper technique's quality is gauged by the specific classifier's accuracy. The wrapper approach typically uses evolutionary or bio-inspired algorithms, in order to guide the search process [26]. It first starts with a population of the solution, also known as a feature subset. Next, the features subset is evaluated via the learning strategy, based on the fitness function. Usually, the existence of a different iteration is used, in order to improve the result. The wrapper approach typically requires high computational costs, along with a higher risk of overfitting, while it then shows a better performance than the filter approach [10]. The most typically used wrapper methods are:

Genetic Algorithm (GA) [27] is a heuristic search algorithm that is inspired by natural evolution and the natural selection process. The main principle of the Genetic Algorithm is to generate a population randomly, while producing offspring with the same inherited characteristics. The algorithm has evolved into three operations: selection, crossover, and mutation. The selection operation chooses the fittest chromosomes, before allowing them to pass to the next generation. Within the crossover operation, two individuals are then selected via the selection operation process. For each individual crossover then, the operation will select a random crossover point and so the two individuals will swap to produce new offspring. Mutations are necessary, in order for there to be a level of diversity maintained in the population.

Artificial Bee Colony (ABC) [28], [29] is a bioinspired evolutionary algorithm, which is essentially inspired by the bees' behaviour when searching for good food sources. The ABC algorithm is based on three groups of bees: employed bees, onlooker bees and scout bees. The employed bees then search for a food source (a solution), as well as sharing information regarding the food source, with onlooker bees then waiting in the hive and dancing. The onlooker bees then select the good food source that has been discovered by the employed bees. The bees that randomly choose these food sources are called scout bees. Any employed bees whose food source does not improve then become scout bees.

Ant Colony Optimization (ACO) [30], [31], is a meta-heuristic algorithm that was originally inspired by ant colonies and the specific foraging behaviour found in them. ACO is based on the fact that the ants can find the shortest path from their colony to the food source and vice versa. Therefore, ACO aims to find an optimal path within the weighted graph. The ants then randomly go searching for any source of food source nearby their colony. When ants finds a food source, they evaluate the quality of the food. Following this, an ant will go back to the colony and leave pheromones (markers) along the way, to guide the other ants to the food source. The other ants can then follow this path with a certain level of probability, as in, a certain amount of pheromones will remain on the path. The path will thus start to get more and more strong as increasing numbers of ants start to find it.

Particle Swarm Optimization (PSO) [32]–[34], is a population-based optimisation technique that is inspired by flocks of birds, fish schooling patterns and the swarming theory. In PSO, each particle represents a candidate solution that then maintains a certain position within the search space. The aim of PSO is for all of the particles to locate the optimum position. Each particle updates its position, by changing its velocity based on its own historical experience, as well as the particle's best performances along with its neighbours, until an optimum position is reached.

Bat Algorithm (BA) [35] is a naturally inspired algorithm, based on micro-bats, which use echolocation behaviour to locate their prey. Bats uses echolocation to measure distances. Therefore, to hunt prey (their solution), the bats fly randomly at a specific velocity and with a fixed frequency towards specific positions, using varying wavelengths and loudness. Their solution is selected from among the best solutions and it is generated through the use of the random walk.

Black Hole Algorithm (BHA) [36] is a population-based optimisation technique, inspired by the behaviour of Black Holes in outer space. The idea behind BHA is that there is a particular region in space that offers no way for any nearby objects to escape its gravitational pull. Therefore, each and every object nearby disappears into into the black hole. BHA typically starts with the initial population of the candidate solutions. For each iteration, the best solution is typically considered to be the black hole, with the others forming normal stars. Following this, the black hole then starts to pull the stars nearer towards it. If a star is swallowed by the black hole, it disappears forever and a new star (solution) is then randomly generated.

Grasshopper Optimization Algorithm (GOA) [37] is a nature-inspired and population-based optimisation algorithm. It is inspired by grasshopper swarms' behaviour. Within this algorithm, grasshoppers' positions represent candidate solutions. The grasshopper's position is defined as its social interaction, the gravity force and the wind advection. The GOA is then used to measure the proximity between two grasshoppers.

Firefly Algorithm (FA) [38], [39] is a natural metaheuristic that is inspired by the global optimisation algorithm, which is in turn inspired by the flashing light patterns and behaviour of firefly insects. Fireflies use their flashing pattern to attract other fireflies (from the opposite sex), as well as their prey. However, the firefly algorithm's development was based on three idealised rules: first, being the assumption that all fireflies are attracted to each other, regardless of their sex. Second, their attractiveness is based on their brightness ability, thus their attractiveness will decrease as the distance between them increases. As a result, it will always be the less bright fireflies that move towards the brighter ones. The brightest ones, meanwhile, will move about randomly. Third, a firefly's brightness is determined or affected by the objective function's form.

Harmony Search algorithm (HSA) [40] is a population-based meta-heuristic algorithm that has been inspired by the search for a perfect state of harmony. When musicians create harmony, they typically first try several possible combinations of music pitches, that are stored in their memory. Three steps are involved in the HAS process. The first step is the Initialisation of the harmony memory (HM), which involves a randomly generated number of solutions. The second step is improvised, so that a new solution can be generated from the HM. Each component of this improved solution is then obtained based on the Harmony Memory Considering Rate (HMCR). HMCR is essentially probability of selecting a solution that comes from the HM. The third step updates the HM, so that if the solution from step two has a better fitness value, then it is naturally replaced by the HM's worst ones.

c: HYBRID OR ENSEMBLE APPROACH

The hybrid approach is built in order to take advantage of both the filter and wrapper approaches. Thus, it combines both the filter approach's computational efficiency with the wrapper approach's high performance [10]. It is based on two stages; the first is for reducing the feature space dimension. Next, the wrapper method is applied so as to select the optimal feature subset. The hybrid model may, however, be less accurate, due to the filter and the wrapper both being done in different steps [41].

The ensemble approach assumes that combination of multiple experts' output is better than the output of any single expert [13]. A single wrapper approach could easily get fantastic results within a single dataset, while it could also perform terribly within another. Therefore, hybridising more than one method results in a lower error rate overall [13].

2) CLASSIFICATION

Classification is a data mining technique that helps assign (predict) the class label that is given to a set of data from within a predefined set of class labels [42]. Classification is a type of supervised learning approach, wherein the class label is defined. Building a classifier is done in two steps: first, the learning phase, where the model (classifier) is constructed according to a set of training data, coupled with a class label. For the second step, the model is used in order to predict the class label for the unseen data, while the accuracy of the classifier is then also measured. The following sections show the more commonly utilised classification methods within the broad field of Microarray data analysis.

a: SUPPORT VECTOR MACHINE SVM [42]–[44]

Is a supervised machine learning algorithm. SVM is centred around searching for a hyperplane that optimally divides the tuples, from one class from another. The hyperplane is found using the *support vector* and the *margin*. The support vector is calculated from the vectors (data points) that define the hyperplane. The margin is the shortest distance between the hyperplane and the nearest point (on two sides). When the data is linearly separable, however, the hyperplane is the line dividing the data into two parts, where each part ends up belonging to a single class. Finding the best hyperplane is achieved via maximising the margin, which is the distance between the nearest data point (called the *support vector*) in either class. Thus, SVM searches for a hyperplane with the highest *maximum marginal hyperplane* (MMH). In a case where the data is not linearly separable, the approach is extended from being a linear SVM to being nonlinear SVM, with the aim then being to find nonlinear hypersurfaces. Real data is typically messy and it is not linearly separable, thus the soft margin classifier is typically used. This allows for some points to be used to violate the hyperplane, thereby allowing for some points to be on the wrong side. The main advantages of using SVM are that it is effective, while it also works well when using high dimensional data. In addition, it works well when the number of features is greater than the number of samples.

b: K NEAREST NEIGHBOUR KNN [45]

K-nearest neighbour is a simple instance-based and non-parametric algorithm for supervised learning. The basic principle of K-NN involves using the similarity metric, while new instances are classified via a process of searching within the training data for the K (most similar instances, like neighbors). Similarity is measured with the distance, such as the Euclidean distance. Following this, the new instances are assigned based on the majority vote of the K (similar neighbours). The main advantage of K-NN is it is simply implemented. It also does not need any distributions on the input data.

c: NAÏVE BAYES (NB) [46], [47]

The Naïve Bayes classifier is a probabilistic classifier, which is based on Bayes' theorem. This classifier is founded on the assumption of *conditional independence*. This means that for every class label, the values of the attributes are essentially conditionally independent of one another. The Naïve Bayes classifier calculates the posterior probability for each class, with the probability for a given attribute value belonging to a class (Maximum A Posteriori (MAP)). Despite this simple assumption, Naïve Bayes has been used to great effect within many situations involving real world data. However, dependencies between attributes do exist. Therefore, the *Bayesian belief networks* have been developed, which allow for class conditional dependencies to exist between the variables.

d: GENETIC PROGRAMING [52]

The Genetic Programing classifier is a one of popular Evolutionary Algorithm (EA), and kind of machine learning. GP inspired by biological evolution and its fundamental procedure, its applied an algorithm that uses random mutation, crossover, a fitness function, and several generations of evolution to discover a user-defined task. GP can be applied to find a functional relationship between features of genes in dataset, to specific class (classification).

e: FUZZY CLASSIFICATION [54]

The Fuzzy Classifcation is the populare algorithim used for classication or grouping into a fuzzy subset whose member function and procuder is identified by the truth value of a fuzzy propositional procedure. Basically, it is used Fuzzy Logic which is starts with and builds on a fuzzy set of user–supplied human language rules.

f: BAGGING CLASSIFIER [59]

The Bagging Classifer is an is an ensemble meta-heuristic algorithim that builds a classifiers each on random subsets of the original dataset. After that aggregate their individual predictions and classification (by voting or by take the avereg) to form a final classification decision.

g: NEURAL NETWORKS (NN) [62]

Neural Networks are also called artificial neural networks, as they simulate the information processing methods that take place in the human brain. Nonlinear data modeling tools are typically used, in order to model the relationship between the different combinations of inputs and outputs. Neural networks typically consist of a group of interconnected artificial neurons. These are then represented in three layers: the input layer, the hidden layer, and the output layer. The input layer represents the records of values, with the number of neurons being equal to the number of features. Several hidden layers may also exist within one single neural network, while the neurons within this layer enact the data transformation to the inputs. The last layer is the output layer. There is a single node (neuron) for each of the class types.

III. HYBRID FEATURE SELECTION METHODS

Feature selection is major research topic in data mining. The main goal of feature selection is to remove the noise and irrelevant features and select the most optimal and informative features. The Microarray data analysis methodology has seen the suggestion of a plethora of feature selection methods. However, this study is aiming to apply the hybrid feature selection approach to select the most informative genes. Two steps are central to the application of the hybrid approach. The first involves a pre-processing step applied to filter off noise and the second involves wrapper techniques that homes in on the subset employing optimum features. Bio-inspired evolutionary methods have widely applied the wrapper approach in feature selection for Microarray data analysis and demonstrate a superior performance [2]. In this study we will review several hybrid and ensemble approaches that employ bio-inspired evolutionary methods as the wrapper method. The performance of such an approach is dependent on two factors; the classification accuracy and the number of selected genes. In this section we will review the most recent works related with gene selection and cancer classification using hybrid approaches that employ bio-inspired evolutionary methods as the wrapper method.

A. HYBRID APPROACH

The majority of gene selection techniques that have been newly created tend to not apply the filter approach on its own [13]. It is also evident from the literature that there is an absence of application of the wrapper approach. This is attributed to the increased risk associated with overfitting and the increased computational cost which are in turn due to the small sample number and increased dimensionality. Instead a hybrid of these methods has been vastly used in the state of the art literature. The hybrid method employs the filter method as a preprocessing step to the wrapper method to avoid its high computational cost and to allow for the selection of the most informative genes with best classification accuracy. The following section will review the latest hybrid methods in their application of evolutionary methods.

1) HYBRID METHOD BASED ON GENETIC ALGORITHM (GA)

Lu *et al.* [48] introduced a novel hybrid feature selection algorithm that combined Mutual Information Maximization (MIM) and the Adaptive Genetic Algorithm (AGA) named MIMAGA-Selection. Initial applications of MIM utilised it as a filter technique to identify genes with high dependency on all other genes. The number of genes obtained from MIM is set to 300. Following this, the AGA was applied. To test the proposed algorithm six multi and binary cancer gene expression datasets were used. The extreme leaning machine (ElM) was selected as a classifier. 30 repeats of the classification process were performed. To compare the accuracy of the MIMAGA-Selection algorithm the authors applied three existing algorithms namely sequential forward selection (SFS), ReliefF and MIM with the ElM classifier on the same dataset with the same target gene number. The result indicates that the accuracy of the MIMAGA-Selection is higher than that of existing feature selection algorithms. Furthermore, the authors classify the selected gene by MIMAGA-Selection using four different classifiers; the back propagation neural network (BP), the support vector machine (SVM), ELM and the regularized extreme learning machine (RELM). All the four classifiers achieve accuracy higher than 80%.

In [49], a hybrid method for feature selection in Microarray data analysis was proposed. The proposed method uses a Genetic Algorithm with Dynamic Parameter setting (GADP) with the X2-test for homogeneity. The chief goal of this technique was to use automated methodology to establish the number of informative genes and to complete this without human interference. First, the BW (between-groups to within-groups sum of square) ratio is used as an initial feature selection method resulting in the selection of 500 genes from the original dataset. Following this the GADP techniques was applied leading to the specification of a gene subset. Then, the X2-test was used to select a specific number of genes generated from (GADP). The SVM classifier was used to evaluate the proposed method. Six cancer datasets were used to measure the performance of the proposed method compared to existing methods. The results illustrated that GADP performed better than existing methods and achieved 100% accuracy in five datasets with a fewer number of genes.

Chuang *et al.* [50] proposed a new hybrid method for gene selection that combined Correlation-based Feature Selection (CFS) and the Taguchi-Genetic Algorithm (TGA). The proposed method was carried out in two phases. The first phase employed the correlation-based feature selection CFS filter method to remove irrelevant features. The subsequent step involved a wrapper approach. This employed the TGA methodology to the features that were generated from the filter step and in doing so identified the best feature subset. The Taguchi-Genetic Algorithm TGA was a hybridization of the Taguchi method and genetic algorithm (GA) where the Taguchi method was inserted in the crossover and mutation operation. It was used as a local search algorithm to

select genes for the crossover operation. The K-nearest neighbour (KNN) classifier was used to evaluate the proposed method in terms of classification accuracy. Eleven cancer datasets with binary and multi class were evaluated. Then, the result of the proposed method was compared to the result obtained by other methods in the literature. The comparison demonstrated that the proposed method achieved the highest classification accuracies in ten datasets where six datasets achieved an accuracy of 100%.

Dashtban and Balafar [51] proposed a novel evolutionary method for gene selection in Microarray data based analysis on the Genetic Algorithm GA and artificial inelegance named Intelligent Dynamic Genetic Algorithm (IDGA). The proposed method was dependent on two phases. In the first phase, Laplacian and Fisher score filter methods were applied to select the top 500 genes. These two methods were applied separately, and their results were examined according to their influence on the proposed method. In the second phase, the IDGA method was applied, which was based on reinforcement learning, and random restart hill climbing. Five Microarray cancer datasets were utilised to assess the suggested methodology. Support Vector machines (SVM), Naïve Bayes (NBY) and K-Nearest Neighbour (KNN) were used as classifiers. The IDGA was performed seven times. The result showed that the proposed method obtained 100% accuracy on four datasets. Moreover, the IDGA with Fisher outperformed the IDGA with Laplacian on four datasets.

Salem *et al.* [52] proposed a feature selection method that integrated Information Gain (IG) and Standard Genetic Algorithm (SGA) referred to as IG/SGA. As a first step the Information gain was applied for feature reduction. Then, GA was employed to select the optimal features determined by IG. The classification was carried out using Genetic Programming GP. The evaluation was carried out in seven cancer Microarray datasets. Findings demonstrated that 100% accuracy was attained in two datasets when using the suggested methodology.

2) A HYBRID METHOD BASED ON ANT COLONY OPTIMIZATION (ACO)

One methodology proposed by Sharbaf *et al.* [53] is the hybrid approach used in the selection of genes for Microarray data analysis. This method combines Cellular Learning Automata and Ant Colony Optimization (CLA-ACO). The proposed approach consisted of three phases. The filter phase was based on the fisher criterion method. The second phase was hybrid method of cellular learning automata and ant colony. A subset with the optimum feature set was the output of this phase. The third phase was to determine the final features subset among the subset resulting from the second phase; the Receiver Operating Characteristics ROC curve was applied and the AUC (area under the curve) was determined. The evaluation of the proposed method was carried out in two stages. Stage one involved the assessment of various ranking methods used to select features. The comparison was on four ranking methods: T-test, information gain, Fisher and Z-score. The Fisher criterion was the best ranking approach. The second stage was the evaluation of the proposed CLA-ACO model. Four cancer datasets with binary and multi-classes were used. The proposed CLACOFS was performed 20 times and the average accuracy was calculated. Three classifiers were compared: SVM, KNN and Naïve Bayes (NB). The results prove that the Naïve Bayes classifier exceeds other classifiers. Moreover, in two datasets, this suggested methodology attained a level of 100% accuracy in two of the datasets.

Vijay and GaneshKumar [54] proposed the Hybrid Stem Cell (HSC) method for designing the Fuzzy classification system for Microarray data analysis. The proposed method utilized Ant Colony optimization (ACO) and novel Adaptive Stem Cell Optimization (ASCO). Informative genes were chosen in the pre-processing step by the application of the Mutual Information (MI) technique. To evaluate the performance of the proposed method five Microarray datasets were examined. The accuracy was compared with other fuzzy based classification systems e.g. Hybrid Colony Algorithm (HCA). The results showed that the proposed method out performed these methods.

3) THE HYBRID METHOD BASED ON THE BAT ALGORITHM (BA)

Dashtbana *et al.* [55] proposed a novel Bio-inspired Multi-objective algorithm for gene selection in Microarray data classification. The proposed algorithm was a multi-objective version of the Bat algorithm (BA) with refined formulations, multi-objective operators and robust local search strategies namely MOBBA-LS. The initial filter methodology employed was the Fisher criterion resulting in the choosing of the upper 500 genes. Then the filtered subset was employed by the proposed MOBBA-LS method. The proposed method was applied to three Microarray cancer datasets. MOBBA-LS was applied 30 times. Throughout the algorithm process, the accuracy of classification was evaluated using the SVM classifier with a 10-fold cross-validation. The accuracy of each subset was evaluated using four classifiers, the support vector machines (SVM), K-Nearest Neighbors (KNN), Naïve Bayes (NBY), and Decision Tree (DT) which was assessed by using Leave-One-Out Cross Validation (LOOCV). The proposed method achieved the highest reported accuracy with a significantly lower number of genes compared to relevant state-of-the-art methods in the Prostate dataset.

4) HYBRID METHOD BASED ON ARTIFICIAL BEE COLONY (ABC)

Aziz *et al.* [56] proposed new hybrid feature selection technique for gene selection in the Microarray data using Independent component analysis (ICA) and Artificial Bee Colony (ABC) termed ICA+ABC. To select the informative genes, ICA initially selected an average of 50 to 180 genes from the original datasets. Following this step, genes that were chosen using ICA were chosen by the ABC algorithm and indicated as a gene subset. The proposed algorithm was

implemented 30 times. The evaluation of ICA+ABC was carried out on six benchmark cancer Microarray datasets. The accuracy of Naïve Bayes classifier was estimated using Leave One Out Cross Validation (LOOCV). Findings indicated that compared to the other current gene selection methodologies in four datasets, the ICA+ABC algorithm achieved the optimum classification accuracy.

Al-shamlan et al. proposed a new hybrid gene selection algorithm that combined minimum redundancy maximum relevance (mRMR) with an artificial bee colony ABC algorithm called mRMR-ABC [57]. Employing mRMR at the pre-processing step acts as a filter method resulting in the decrease in the number of non-essential genes. Thus, mRMR was combined with SVM classifier to select a set of genes that attained 100% accuracy. Following this step, the major informative genes were indicated by the application of the ABC algorithm on the data generated from the mRMR dataset. Finally, the Support Vector Machine (SVM) classifier was used to measure the efficiency of the proposed method. The evaluation of mRMR-ABC is carried out on six binary and multiclass gene expression Microarray cancer datasets. Additionally, the proposed algorithm was compared with (mRMR-GA) and (mRMR-PSO) as well as recently published gene selection methods. The results indicated that the mRMR-ABC methodology achieved 100% accuracy in five datasets.

5) HYBRID METHOD BASED ON PARTICLE SWARM OPTIMIZATION (PSO)

Moradi and Gholampour [26] proposed a novel hybrid feature selection algorithm based on Particle Swarm Optimization (PSO) and the local search strategy termed HPSO-LS. The proposed method was hybridized with the local search in PSO. The aim of the local search was to use features correlation data to direct the PSO search procedure in selection of specific features. The feature selection in HPSO-LS was carried out in seven steps. In the first step, the size of the final subset was determined automatically using a probabilistic random function. The second step utilized the correlation values to split the features into two groups based on features similarity. From step three to seven, the hybridization of the POS with the local search was carried out. To evaluate the performance of the proposed method twelve cancer datasets were used. In the experiments the k-nearest neighbour (k-NN) was used. The accuracy of the proposed method was compared to previously proposed wrapper-based methods.

Jain *et al.* [58] proposed a hybrid method gene selection in Microarray data. The method combined Correlation-based Feature Selection (CFS) with improved-Binary Particle Swarm Optimization (iBPSO) termed CFS-iBPSO. The proposed method operated on a hybrid framework of two phases; a filter phase and a wrapper phase. The filter phase employed multivariate filter Correlation-based Feature Selection CFS. The next phase applied the wrapper method iBPSO that selected an optimal subset from the gene subset resulting from the filter phase. The Naive–Bayes classifier (NB) with 10-fold-cross-validation was used to evaluate the proposed method. Eleven Microarray datasets were tested. The result was compared with seven existing feature selection methods. CFS-iBPSO achieved the highest accuracy in 10 data sets and 100% accuracy in seven datasets.

6) HYBRID METHOD BASED ON BLACK HOLE ALGORITHM (BH)

Pashaei *et al.* [59] proposed a novel method for gene selection in Microarray data based on the Binary Black Hole Algorithm (BBHA) and Random Forest Ranking (RFR). RFR-BBHA-Bagging ordered the genes by employing RFR as a filter method. Then the top 500 ranked genes were grouped to a new gene subset that will be input to BBAH. Following this step, formative genes were chosen from the subset generated by the preceding step using the Black Hole Algorithm. To evaluate the proposed method the Bagging classifier with 10-fold cross validation was applied on four Microarray cancer data. The proposed method was applied 100 times. Finally, the RFR-BBHA-Bagging was compared to seven known classifiers; the result indicated that the proposed method achieved highest accuracy in two datasets.

7) HYBRID METHOD BASED ON BIOGEOGRAPHY ALGORITHM

Li and Yin [60] proposed a Multi-Objective Binary Biogeography (MOBBBO) based gene selection method. As a first step, the Fisher-Markov selector was applied to select the top 60 genes. Then, MOBBBO was applied to select the most informative gene subset with the binary migration method and a binary mutation model. Next, the suggested methodology was assessed using the SVM classifier. Ten Microarray datasets were tested, and the result was compared with three PSO based methods. MOBBBO+SVM achieved the highest accuracy in nine datasets where three of them obtained 100%.

8) HYBRID METHOD BASED ON HARMONY SEARCH ALGORITHM (HSA)

Shreem *et al.* [61] proposed a hybrid filter-wrapper method for gene selection in Microarray data that combined Symmetrical Uncertainty (SU) with a Harmony Search Algorithm (HSA) referred to as SU-HSA. The methodology was completed over two steps. In the initial step non-essential and superfluous gene data was eliminated using the SU method. The second stage used HAS as a wrapper approach to select the most informative genes. Two classifiers were employed to evaluate the proposed method; IB1 and Naive Bayes (NB). The experiment was carried out in 10 Microarray datasets and it was repeated 10 times. Compared to the state-of-the-art gene selection method, the proposed method achieves the highest accuracy in five data sets with 100% accuracy in four of them.

9) HYBRID METHOD BASED ON GRASSHOPPER OPTIMIZATION ALGORITHM GOA

Tumuluru and Ravi [62] proposed the Grasshopper Optimization Algorithm-based Deep belief Neural Networks (GOA-based DBN) for feature selection in Microarray data.

The initial data management involved two stages; a first pre-processing stage and a second gene selection stage. In the pre-processing step Logarithmic transformation was used to reduce the skewness of the data. Bhattacharya distance was applied in the gene selection step to select relevant features and remove redundant features. Following this step, the proposed method GOA-based DBN was applied to select the optimal gene subset. To evaluate the proposed method two Microarray datasets were tested and the result was compared with three existing methods. The proposed method achieves the highest accuracy among the compared methods.

B. ENSEMBLE APPROACH

Ensemble approach is built on the assumption that: combining the output of multiple experts is better than the output of a single expert [13]. This approach is recently applied to Microarray gene selection and Microarray data classification problems. Any single wrapper approach may show completely contradictory results; attaining excellent results in certain datasets while failing in others. Thus, hybridizing more than one method will result in a lower error rate such as hybridizing the genetic algorithm with the artificial bee colony [13]. The following section with discuss various algorithms that utilize the ensemble approach with natural inspired evolutionary methods.

Authors in [63] present a metaheuristic framework for gene selection using Harmony Search (HS) with Genetic Algorithm (GA). The proposed model was dependent on two phases. In the first phase the HS method was embedded in the GA process where randomly selected solutions were ranked based on their fitness value. In the second phase, GA selected the top ranked solution to generate the offspring (new solution). Following this step, the fitness of the new solution was calculated using the SVM classifier. The proposed model was compared using seven existing feature selection methods and it achieved a lower error rate with higher accuracy among them.

Chuang *et al.* [64] proposed A hybrid method of Binary Particle Swarm Optimization (BPSO) and a Combat Genetic Algorithm (CGA) for gene selection Microarray data. The compact GA is embedded in BPSO and acts as a local optimizer for each generation. Thus, compact GA was applied to generated particles, for reproduction, crossover and mutation. The K-nearest neighbour (K-NN) was used as a classifier. Ten Microarray datasets were examined to assess the suggested methodology. This was subsequently compared to four methods reported in the literature. The result demonstrated that the proposed method achieved the lowest error rate in nine datasets.

Djellali *et al.* [65] presented and compared two hybrid methods. The first method was a Fast Correlation based Filter FCBF with Genetic Algorithm GA (FCBF-GA); the second method was FCBF with Particle Swarm Optimization PSO (FCBF-POS). The first stage in the proposed methods was to apply FCBF as a filter technique. The findings were subsequently applied as input for GA and POS. The evaluation of the proposed methods was carried out using four Microarray datasets and Support Vector Machine SVM as classifier. The experiment result showed that the performance of the FCBF-POS was better than FCBF-GA in terms of the accuracy and the number of selected genes.

Al-shamlan et al. proposed a new hybrid gene selection method, Genetic Bee Colony (GBC) that combined the genetic algorithm (GA) with the Artificial Bee Colony (ABC) algorithm [66]. In the proposed GBC, GA operations were hybridized with ABC in the onlooker bee (crossover operation) phase and in the scout bee phase (mutation operation). The mRMR filter methodology was employed together with the SVM classifier to choose a gene set that included genes that attained 100% accuracy. The performance evaluation of the proposed algorithm was carried out on six binary and multi-class Microarray datasets. The accuracy of the classification was determined using the SVM classifier/ the evaluation experiment was carried out 30 times for each dataset. The proposed algorithm was compared with (mRMR-ABC), (mRMR-GA) and (mRMR-PSO) in addition to recently published gene selection methods.

IV. ANALYSIS AND DISCUSSION

Table 1 shows the performance of various hybrid approaches for gene selection and cancer classification in Microarray data. Considering the result in table 1, it obviously can be seen that the hybrid method shows superior performance in terms of high accuracy and small number of selected genes. This is because the hybrid algorithm deal perfectly with high dimensionality and over -fitting problems by applying filter approach first as preprocessing step in order to reduce the dimensionality of Microarray gene expression profile.

Microarray cancer classification suffering from over fitting problem. This is due to curse of dimensionality: small number of sample and large number of features. Thus, to avoid it we highly recommend applying hybrid approaches in order to adopt filtering before start the classification process. Also, (supervised) machine learning will be more effected with over fitting problem with applying Leave–One–Out Cross Validation (LOOCV) approach for experiment to hold out part of the available data as a test set. A test set should still be held out for final evaluation, but the validation set is no longer needed when doing CV.

Moreover, in this research problem, the parameters fitting are not an easy process. Because, it will be affect the classification accuracy. In cancer gene expression data classification the parameter fitting is depending on the gene expression datasets and the applied feature selection and classification and methods. Thus, different dataset have different parameters value, which are not fit for all algorithms. Based on our research, we noted that the most researchers adjust the parameters manually. They try different values until they reach the accepted results.

It is evident from Table 1 that the Mutual Information Maximization (MIM) and Genetic Algorithm (GA)

TABLE 1. Comparison table.

Reference	Filter Method	Wrapper method	Classifier	Datasets used	Classification accuracy	No. of selected genes
Huijuan Lu et al. [48]	Mutual Information Maximization (MIM)	Genetic Algorithm (GA)	SVM	*Colon* [68]	83.41	202
C.-P. Lee and Y. Leu [49]	X^2-test	Genetic Algorithm (GA)	SVM	*Colon* [68]	100	8
				DLBCL [69]	100	6
				SRBCT [70]	100	8
				Leukemia1 [71]	100	5
Chang et al. [50]	Correlation-based Feature Selection (CFS)	Genetic Algorithm (GA)	KNN	*SRBCT* [70]	100	29
				Prostate [67]	99.22	24
				Lung [67]	98.42	195
Dashtban and Balafar [51]	Laplacian and Fisher score	Genetic Algorithm (GA)	SVM	*SRBCT* [70]	100	18
				Leukemia1 [71]	100	15
				Prostate [72]	96.3	14
				Breast [73]	100	2
				DLBCL [69]	100	9
			KNN	*SRBCT* [70]	91.6	NAN
				Leukemia1 [71]	97.2	NAN
				Prostate [72]	95.6	NAN
				Breast [73]	95.5	NAN
				DLBCL [69]	97.9	NAN
			NB	*SRBCT* [70]	98.2	NAN
				Leukemia1 [71]	93.1	NAN
				Prostate [72]	93.4	NAN
				Breast [73]	100	NAN
				DLBCL [69]	95.8	NAN
Salem et al. [52]	Information Gain (IG)	Genetic Algorithm (GA)	GP	*Colon* [68]	85.48	60
				Leukemia1 [71]	97.06	3
				Lung [74]	100	9
				Prostate [72]	100	26
Sharbaf et al. [53]	Fisher Criterion	Ant Colony Optimization (ACO)	SVM	*Leukemia1* [71]	95.95	3
				Prostate [72]	98.35	14

TABLE 1. *(Continued.)* **Comparison table.**

Reference	Filter Method	Wrapper method	Classifier	Datasets used	Classification accuracy	No. of selected genes
			KNN	*Leukemia1* [71]	94.30	3
				Prostate [72]	99.25	15
			NB	*Leukemia1* [71]	95.95	4
				Prostate [72]	99.40	10
Vijay and GaneshKumar [54]	Mutual Information (MI)	Ant Colony Optimization (ACO)	Fuzzy classification	*Colon* [68]	100	NAN
				Leukemia1 [71]	100	NAN
				Prostate [67]	90.85	NAN
Dashtbana et al. [55]	Fisher criterion	Bat algorithm (BA)	SVM	*SRBCT* [70]	85	6
				Prostate [72]	94.1	6
			KNN	*SRBCT* [70]	100	6
				Prostate [72]	97.1	6
			NB	*SRBCT* [70]	100	6
				Prostate [72]	97.1	6
Aziz et al [56]	Independent component analysis (ICA)	Artificial Bee Colony (ABC)	NB	*Colon* [68]	98.14	16
				Leukemia1 [71]	98.68	12
				Leukema2 [75]	97.33	15
				Lung [73]	92.45	24
Al-shamlan et al. [57]	minimum redundancy maximum relevance (mRMR)	Artificial Bee Colony (ABC)	SVM	*Colon* [68]	96.77	15
				SRBCT [70]	100	10
				Leukemia1 [71]	100	14
				Leukema2 [75]	100	20
				Lung [74]	100	8
				Lymphoma [76]	100	5
Jain et al. [58]	Correlation-based Feature Selection (CFS)	Particle Swarm Optimization (PSO)	NB	*Colon* [68]	94.89	4
				SRBCT [70]	100	34
				Leukemia1 [71]	100	4
				Leukema2 [75]	100	6
				Lymphoma [76]	100	24
				MILL [77]	100	30
				Breast [73]	100	10
P. Moradi and M.	probabilistic random	Particle Swarm	KNN	*Colon* [68]	84.38	60

TABLE 1. *(Continued.)*. **Comparison table.**

Reference	Filter Method	Wrapper method	Classifier	Datasets used	Classification accuracy	No. of selected genes
Gholampour[26]	function	Optimization (PSO)		*Leukemia1* [71]	89.28	100
				Lymphoma [76]	87.71	50
Pashaei et al. [59]	Random Forest Ranking (RFR)	Black Hole Algorithm (BHA)	Bagging Classifier	*Colon* [68]	91.93	3
				MILL [77]	98.61	5
Li and Yin [60]	Fisher-Markov selector	Biogeography algorithm	SVM	*SRBCT* [70]	100	6
				Prostate [67]	98.3	12
				Lung [67]	98.4	16
Shreem et al. [61]	Symmetrical Uncertainty (SU)	Harmony Search Algorithm (HSA)	NB	*Colon* [68]	87.53	9
				SRBCT [70]	99.89	37
				Leukemia1 [71]	100	26
				Leukema2 [75]	100	24
				Lymphoma [76]	100	10
				MILL [77]	98.97	10
Tumuluru and Ravi [62]	Logarithmic transformation	Grasshopper Optimization Algorithm	NN	*Colon* [68]	95	NAN
				Leukemia1 [71]	94	NAN
Djellali et al. [65]	Fast Correlation based Filter FCBF	Particle Swarm Optimization PSO AND Genetic Algorithm GA	*SVM*	*Colon* [68]	96.30	1000
				DLBCL [69]	100	3204
Al-shamlan et al. [66]	minimum redundancy maximum relevance (mRMR)	Genetic Bee Colony (GBC) AND Genetic algorithm (GA)	SVM	*Colon* [68]	98.38	10
				SRBCT [70]	100	6
				Leukemia1 [71]	100	4
				Leukema2 [75]	100	8
				Lung [74]	100	4
				Lymphoma [76]	100	4

technique [48] generated the largest number of selected genes and resulted in the poorest classification accuracy. In the *Lung* [67] datasets Correlation-based Feature Selection (CFS) with Genetic Algorithm (GA) method [50] selected the highest number of genes. While the IG/SGA [52] method attained 100% accuracy in the *Lung* [74] dataset but with the highest number of selected genes. The proposed Fast Correlation based Filter FCBF with Particle Swarm Optimization PSO AND Genetic Algorithm GA method [65] selects the highest number of genes in the DLBCL dataset 3204 compared to other methods as shown in table 1. In addition, the greatest level of accuracy was generated in the colon dataset using X^2-test with GADP Genetic Algorithm (GA) [49] and by utilising a small quantity of genes. In the SRBCT dataset the proposed minimum redundancy maximum relevance (mRMR) with Genetic

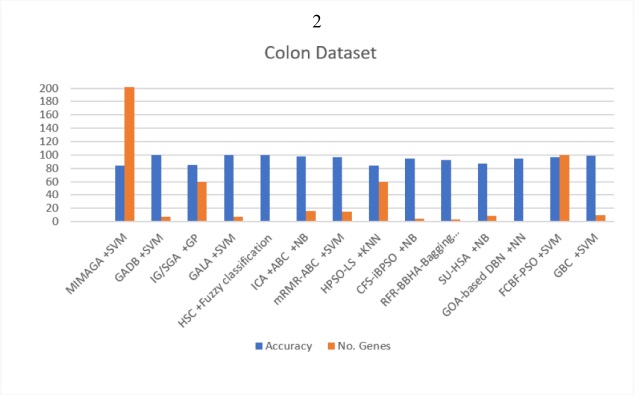

FIGURE 2. Performance result on colon dataset.

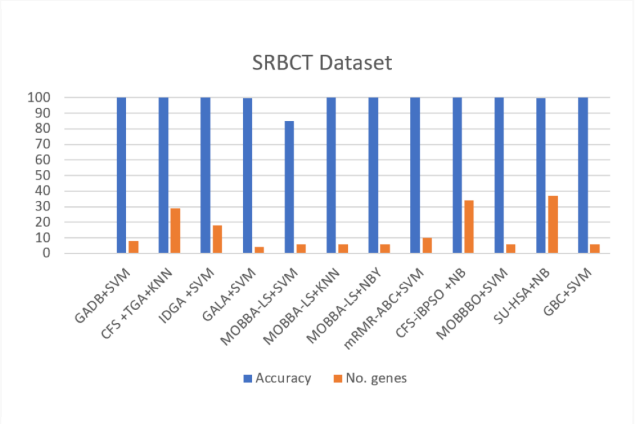

FIGURE 3. Performance result on SRBCT dataset.

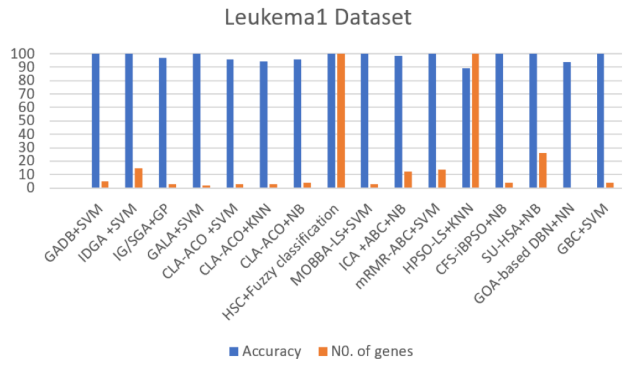

FIGURE 4. Performance result on leukema1 dataset.

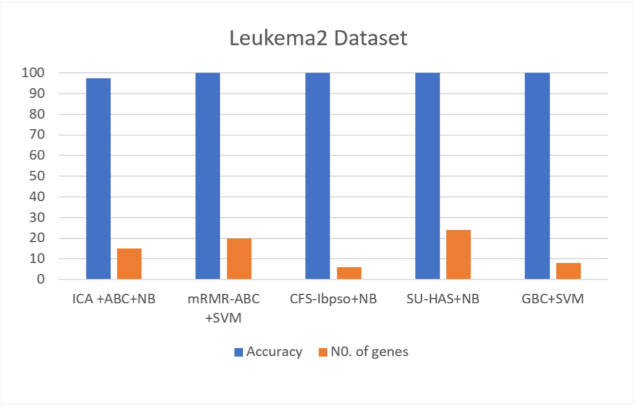

FIGURE 5. Performance result on leukema2 dataset.

Bee Colony (GBC) AND Genetic algorithm (GA) [66] algorithm has the highest accuracy with the smallest number of genes. Moreover, the Independent component analysis (ICA) with Particle Swarm Optimization (PSO) methodology [26] achieved the least score for accuracy using the greatest quantity of selected genes.

The genetic algorithm is the most applied wrapper method in the literature. Among all other applied methods genetic algorithm achieves the highest accuracy with relatively small numbers of selected genes. As shown in table 1 GA obtained 100% accuracy of the most data sets in 5 out of 6 reported algorithms. All algorithms that were reported demonstrated good performance in terms of Ant colony optimization. As presented in table 1 ACO obtained an accuracy of more than 90% on all methods employing a small number of genes less than 15 genes were selected. The artificial bee colony attained high classification accuracy with more than 98% and less than 15 selected genes in all reported algorithms. However, ACO achieved 100% accuracy when combined with the SVM classifier. Particle Swarm Optimization attained a high accuracy of 100% in two out of 4 reported methods. However, it selected a high number of genes compared to the other wrapper methods. As previously demonstrated, the Firefly algorithm has not been used as a wrapper method in gene selection for Microarray data classification.

As presented in figure 2, eight out of fourteen reported method achieve accuracy more than 95% when applied to colon dataset. Where the algorithm that get the highest accuracy select the smallest number of gene. However, the algorithms that has the lowest accuracy select the highest number of genes.

All the twelve reported algorithms get an accuracy more than 99.8% when applied to SRBCT dataset as presented on figure 3. Worth to mention that most of the algorithms select low number of informative genes.

Eight of fifteen applied method on leukemia1 dataset achieve 100% accuracy as reported in figure 4. However, three of them get an accuracy less than 95%. Where, two of the applied algorithm select high number of gene (100 genes)

As reported figure 5 four out of five applied algorithms on leukemia2 dataset achieved 100% of accuracy. While the applied algorithms select relatively high number of selected genes.

As presented on figure 6 one out of ten applied algorithms to Prostate dataset achieved 100% accuracy where in the other nine the accuracy ranged from 93% to 97%. However, most of the applied algorithms selected high number of genes Four out five applied algorithms to Lymphoma dataset got accuracy of 100% as presented in figure 7. On the other hand,

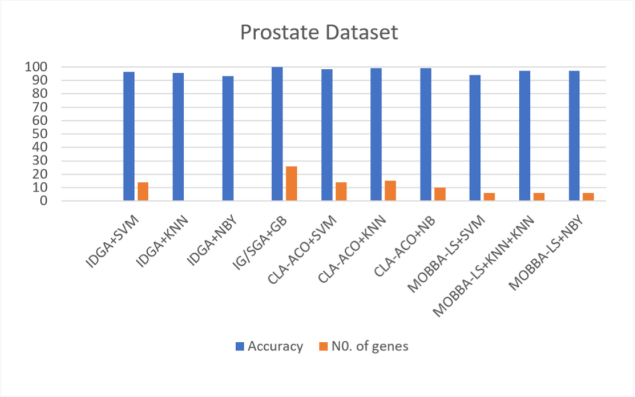

FIGURE 6. Performance result on prostate dataset.

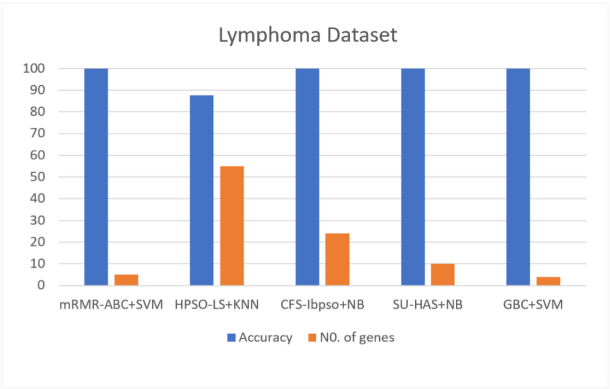

FIGURE 7. Performance result on lymphoma dataset.

the algorithm that got the lowest accuracy selected the lower number of genes.

Given the above we can clearly notice that the lower the accuracy the more the selected genes. Hence, the algorithms that achieve high accuracy select low number of genes.

V. CONCLUSION

Microarray data analysis provides valuable results which contribute towards solving gene expression profile problems. One the most important applications of Microarray data analysis is cancer classification. Classification is challenging due to the high dimensionality found in a small sample size of gene expression data. The most practical method to overcome these challenges is therefore a feature selection technique. Many hybrid algorithms that employ bio-inspired method as wrapper technique have been used for gene selection and cancer classification in Microarray data analysis. In order to review and compare these algorithms we conducted this study. We can conclude that the genetic algorithm GA is the most applied wrapper method in the literature. Among all other applied methods genetic algorithm achieves the highest accuracy with relatively small numbers of selected genes.

As previously demonstrated, the Firefly algorithm has not been used as a wrapper method in gene selection for Microarray data classification. Therefore, as future work we aim to apply a hybrid gene selection algorithm based on a filter and wrapper approach to identify the most informative genes for cancer classification. A Firefly wrapper method will employ to identify the optimal gene subset

REFERENCES

[1] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," *IEEE Trans. Comput.*, vol. C-26, no. 9, pp. 917–922, Sep. 1977.
[2] H. Alshamlan, G. Badr, and Y. Alohali, "A comparative study of cancer classification methods using microarray gene expression profile," in *Proc. 1st Int. Conf. Adv. Data Inf. Eng. (DaEng)*. Singapore: Springer, 2014, pp. 389–398.
[3] M. M. Babu, "Introduction to microarray data analysis," in *Computational Genomics: Theory and Application*. 2004, pp. 225–249.
[4] J. Read and S. Brenner, "Microarray technology," in *Encyclopedia of Genetics*. New York, NY, USA: Academic, 2001, p. 1191.
[5] A. Perez-Diez, A. Morgun, and N. Shulzhenko, *Microarrays for Cancer Diagnosis and Classification*. 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, Jun. 2009.
[7] C. Gunavathi, K. Premalatha, and K. Sivasubramanian, "A survey on feature selection methods in microarray gene expression data for cancer classification," *Res. J. Pharmacy Technol.*, vol. 10, no. 5, pp. 1395–1401, 2017.
[8] A. L. Tarca, R. Romero, and S. Draghici, "Analysis of microarray experiments of gene expression profiling," *Amer. J. Obstetrics Gynecol.*, vol. 195, no. 2, pp. 373–388, Aug. 2006.
[9] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "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, Jul./Aug. 2012.
[10] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007.
[11] J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Comput. Appl.*, vol. 24, no. 1, pp. 175–186, Jan. 2014.
[12] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Adv. Bioinf.*, vol. 2015, May 2015, Art. no. 198363. Accessed: Mar. 28, 2018. [Online]. Available: https://www.hindawi.com/journals/abi/2015/198363/
[13] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Inf. Sci.*, vol. 282, pp. 111–135, Mar. 2014.
[14] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[15] M. A. Hall, "Correlation-based feature selection for machine learning," Univ. Waikato, Hamilton, New Zealand, Tech. Rep. 19, Apr. 1999.
[16] L. S. Kao and C. E. 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, Jan. 2008.
[17] Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," Feb. 2012, *arXiv:1202.3725*. [Online]. Available: https://arxiv.org/abs/1202.3725
[18] Q. Cheng, H. Zhou, and J. Cheng, "The Fisher-Markov selector: Fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1217–1233, Jun. 2011.
[19] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in *Proc. 18th Int. Conf. Neural Inf. Process. Syst.*, 2016, p. 8.
[20] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Comput.*, vol. 9, no. 7, pp. 1483–1492, Jul. 1997.
[21] C.-H. Zheng, D.-S. Huang, and L. Shang, "Feature selection in independent component subspace for microarray data classification," *Neurocomputing*, vol. 69, nos. 16–18, pp. 2407–2410, Oct. 2006.
[22] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[23] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Mach. Learn.*, vol. 6, no. 1, pp. 37–66, Jan. 1991.
[24] G. Xuan, X. Zhu, P. Chai, Z. Zhang, Y. Q. Shi, and D. Fu, "Feature selection based on the Bhattacharyya distance," in *Proc. 18th Int. Conf. Pattern Recognit.*, vol. 3, Washington, DC, USA, Aug. 2006, p. 957.
[25] C. C. Reyes-Aldasoro and A. Bhalerao, "The Bhattacharyya space for feature selection and its application to texture segmentation," *Pattern Recognit.*, vol. 39, no. 5, pp. 812–826, May 2006.
[26] P. Moradi and 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, Jun. 2016.
[27] J. McCall, "Genetic algorithms for modelling and optimisation," *J. Comput. Appl. Math.*, vol. 184, no. 1, pp. 205–222, Dec. 2005.
[28] D. Karaboga, "An idea based on honey bee swarm for numerical optimization," Dept. Comput. Eng., Erciyes Univ., Kayseri, Turkey, Tech. Rep. TR06, 2005, p. 10.
[29] Y. Xu, P. Fan, and L. Yuan, "A simple and efficient artificial bee colony algorithm," *Math. Problems Eng.*, vol. 2013, Dec. 2012, Art. no. 526315. Accessed: Apr. 4, 2018. [Online]. Available: https://www.hindawi.com/journals/mpe/2013/526315/
[30] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant system: Optimization by a colony of cooperating agents," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 26, no. 1, pp. 29–41, Feb. 1996.
[31] M. Dorigo and C. Blum, "Ant colony optimization theory: A survey," *Theor. Comput. Sci.*, vol. 344, nos. 2–3, pp. 243–278, Nov. 2005.
[32] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Netw.*, vol. 4, Nov./Dec. 1995, pp. 1942–1948.
[33] R. Poli, J. Kennedy, and T. Blackwell, "Particle swarm optimization: An overview," *Swarm Intell.*, vol. 1, no. 1, pp. 33–57, Oct. 2007.
[34] J. Brownlee, *Clever Algorithms: Nature-Inspired Programming Recipes*. Austalia, 2011.
[35] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in *Nature Inspired Cooperative Strategies for Optimization (NICSO 2010)*. Berlin, Germany: Springer, 2010, pp. 65–74.
[36] A. Hatamlou, "Black hole: A new heuristic optimization approach for data clustering," *Inf. Sci.*, vol. 222, pp. 175–184, Feb. 2013.
[37] S. Saremi, S. Mirjalili, and A. Lewis, "Grasshopper optimisation algorithm: Theory and application," *Adv. Eng. Softw.*, vol. 105, pp. 30–47, Mar. 2017.
[38] X.-S. Yang, "Firefly algorithms for multimodal optimization," in *Proc. Int. Symp. Stochastic Algorithms*. Berlin, Germany: Springer, 2009, pp. 169–178.
[39] X.-S. Yang, *Nature-Inspired Metaheuristic Algorithms*. London, U.K.: Luniver Press, 2008.
[40] Z. W. Geem, J. H. Kim, and G. V. Loganathan, "A new heuristic optimization algorithm: Harmony search," *J. Simul.*, vol. 76, no. 2, pp. 60–68, Feb. 2001.
[41] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *Proc. 38th Int. Conv. Inf. Commun. Technol., Electron. Microelectron. (MIPRO)*, May 2015, pp. 1200–1205.
[42] J. Han, J. Pei, and M. Kamber, *Data Mining: Concepts and Techniques*. Amsterdam, The Netherlands: Elsevier, 2011.
[43] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.
[44] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop Comput. Learn. Theory*, New York, NY, USA, 1992, pp. 144–152.
[45] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.
[46] S. Taheri and M. Mammadov, "Learning the naive Bayes classifier with optimization models," *Int. J. Appl. Math. Comput. Sci.*, vol. 23, no. 4, pp. 787–795, Dec. 2013.
[47] L. I. Kuncheva, "On the optimality of Naïve Bayes with dependent binary features," *Pattern Recognit. Lett.*, vol. 27, no. 7, pp. 830–837, May 2006.
[48] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, "A hybrid feature selection algorithm for gene expression data classification," *Neurocomputing*, vol. 256, pp. 56–62, Sep. 2017.
[49] C.-P. Lee and Y. Leu, "A novel hybrid feature selection method for microarray data analysis," *Appl. Soft Comput.*, vol. 11, no. 1, pp. 208–213, Jan. 2011.
[50] L.-Y. Chuang, C.-H. Yang, K.-C. Wu, and C.-H. Yang, "A hybrid feature selection method for DNA microarray data," *Comput. Biol. Med.*, vol. 41, no. 4, pp. 228–237, Apr. 2011.
[51] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol. 109, no. 2, pp. 91–107, Mar. 2017.
[52] H. Salem, G. Attiya, and N. El-Fishawy, "Classification of human cancer diseases by gene expression profiles," *Appl. Soft Comput.*, vol. 50, pp. 124–134, Jan. 2017.
[53] F. V. Sharbaf, S. Mosafer, and M. H. Moattar, "A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization," *Genomics*, vol. 107, no. 6, pp. 231–238, Jun. 2016.
[54] S. A. A. Vijay and P. G. Kumar, "Fuzzy expert system based on a novel hybrid stem cell (HSC) algorithm for classification of micro array data," *J. Med. Syst.*, vol. 42, no. 4, p. 61, Apr. 2018.
[55] M. Dashtban, M. Balafar, and P. Suravajhala, "Gene selection for tumor classification using a novel bio-inspired multi-objective approach," *Genomics*, vol. 110, no. 1, pp. 10–17, Jan. 2018.
[56] R. Aziz, C. K. Verma, and N. Srivastava, "A novel approach for dimension reduction of microarray," *Comput. Biol. Chem.*, vol. 71, pp. 161–169, Dec. 2017.
[57] H. Alshamlan, G. Badr, and Y. Alohali, "mRMR-ABC: A hybrid gene selection algorithm for cancer classification using microarray gene expression profiling," *Biomed. Res. Int.*, vol. 2015, Mar. 2015, Art. no. 604910. Accessed: Feb. 19, 2018. [Online]. Available: https://www.hindawi.com/journals/bmri/2015/604910/abs/
[58] I. Jain, V. K. Jain, and R. Jain, "Correlation feature selection based improved-binary particle swarm optimization for gene selection and cancer classification," *Appl. Soft Comput.*, vol. 62, pp. 203–215, Jan. 2018.
[59] E. Pashaei, M. Ozen, and N. Aydin, "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)*, Feb. 2016, pp. 308–311.
[60] X. Li and M. Yin, "Multiobjective binary biogeography based optimization for feature selection using gene expression data," *IEEE Trans. Nanobiosci.*, vol. 12, no. 4, pp. 343–353, Dec. 2013.
[61] S. S. Shreem, S. Abdullah, and M. Z. A. Nazri, "Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm," *Int. J. Syst. Sci.*, vol. 47, no. 6, pp. 1312–1329, Apr. 2016.
[62] P. Tumuluru and 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.
[63] K. Das, D. Mishra, and K. Shaw, "A metaheuristic optimization framework for informative gene selection," *Inform. Med. Unlocked*, vol. 4, pp. 10–20, Jan. 2016.
[64] L.-Y. Chuang, C.-H. Yang, J.-C. Li, and C.-H. Yang, "A hybrid BPSO-CGA approach for gene selection and classification of microarray data," *J. Comput. Biol.*, vol. 19, no. 1, pp. 68–82, Jan. 2011.
[65] H. Djellali, S. Guessoum, N. Ghoualmi-Zine, and S. Layachi, "Fast correlation based filter combined with genetic algorithm and particle swarm on feature selection," in *Proc. 5th Int. Conf. Elect. Eng.-Boumerdes (ICEE-B)*, Oct. 2017, pp. 1–6.
[66] H. M. Alshamlan, G. H. Badr, and Y. A. Alohali, "Genetic bee colony (GBC) algorithm: A new gene selection method for microarray cancer classification," *Comput. Biol. Chem.*, vol. 56, pp. 49–60, Jun. 2015.
[67] *GEMS: Gene Expression Model Selector*. Accessed: Mar. 16, 2018. [Online]. Available: http://www.gems-system.org/
[68] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," *Proc. Nat. Acad. Sci. USA*, vol. 96, no. 12, pp. 6745–6750, Jan. 1999.
[69] *UCI Machine Learning Repository: Data Sets*. Accessed: Mar. 16, 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets.html
[70] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Med.*, vol. 7, no. 6, pp. 673–679, Jun. 2001.
[71] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.
[72] *Gene Expression Correlates of Clinical Prostate Cancer Behavior: Cancer Cell*. Accessed: Mar. 16, 2018. [Online]. Available: http://www.cell.com/cancer-cell/fulltext/S1535-6108(02)00030-2

[73] *Microarray Datasets*. Accessed: Mar. 6, 2018. [Online]. Available: http://csse.szu.edu.cn/staff/zhuzx/Datasets.html
[74] D. G. Beer, S. L. R. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas, M. L. Lizyness, R. Kuick, S. Hayasaka, J. M. G. Taylor, M. D. Iannettoni, M. B. Orringer, and S. Hanash, "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Med.*, vol. 8, no. 8, pp. 816–824, Aug. 2002.
[75] S. A. Armstrong, J. E. Staunton, L. B. Silverman, R. Pieters, M. L. den Boer, M. D. Minden, S. E. Sallan, E. S. Lander, T. R. Golub, and S. J. Korsmeyer, "MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia," *Nature Genet.*, vol. 30, no. 1, pp. 41–47, Jan. 2002.
[76] A. A. Alizadeh *et al.*, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, no. 6769, pp. 503–511, Feb. 2000.
[77] *MLDATA:: Repository :: :: Leukemia MLL*. Accessed: Mar. 16, 2018. [Online]. Available: http://mldata.org/repository/data/viewslug/leukemia-mll/
[78] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics* vol. 17, no. 6, pp. 520–525, 2001.
[79] I. Lobo, "Same genetic mutation, different genetic disease phenotype," *Nature Educ.*, vol. 1, no. 1, pp. 64–71, 2008.
[80] P. Yang, B. B. Zhou, Z. Zhang, and A. Y. Zomaya, "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.

NADA ALMUGREN received the bachelor's and master's degree in information technology from King Saud University, in 2015 and 2018, respectively. She has been a member of the Research Faculty, since 2018. She is currently an Academic Researcher with King Abdulaziz City for Science and Technology.

Her research interests include computer science are in the areas of bioinformatics, microarray data analysis, data mining, artificial intelligence, and solving gene expression profile problems which aim is to identify the most informative genes contributing to cancer diagnosis.

HALA ALSHAMLAN received the Ph.D. degree in computer science from King Saud University, in 2015. From 2016 to 2017, she was a Research Fellow with the Kamm Lab, Mechanical Engineering Department, Massachusetts Institute of Technology (MIT), Cambridge, USA. In 2018, she was leading of Data Science for Global Health track at MIT Hacking Medicine, Riyadh, Saudi Arabia.

She is currently an Assistant Professor with the Information Technology Department, College of Computer and Information Sciences, King Saud University (KSU). She is interested in data science and big data analytics. She developed many novel algorithms that discover the cancer biomarker from genomic data. All these algorithms have been published on high impact journals. Her research interests include bioinformatics especially on how apply artificial intelligence techniques and machine learning approaches to analyze biological data.