

Received 25 May 2022, accepted 14 June 2022, date of publication 22 June 2022, date of current version 11 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3185226

# A Comprehensive Survey of Recent Hybrid Feature Selection Methods in Cancer Microarray Gene Expression Data

HALAH ALMAZRUA AND HALA ALSHAMLAN<sup>ID</sup>

Department of Information Technology, King Saud University, Riyadh 11362, Saudi Arabia

Corresponding author: Hala Alshamlan (halshamlan@ksu.edu.sa)

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** In the diagnosis and treatment of cancer, cancer classification is a vital issue. Gene selection is much needed to solve the high dimensionality issue in microarray data, small sample size, and noisy. The best way to classify cancer is to select those genes that hold the most informative ones, and this process contributes significantly to the classification performance of microarrays. In this survey, we comprehensively studied hybrid selection methods proposed since 2017, that may be used for comparison to several other algorithms proposed for gene selection in cancer classification in the past and looked to see if there are any challenges future authors that need to be discussed.

**INDEX TERMS** Bio-inspired, meta-heuristic, swarm intelligence, biomarker discovery, cancer classification, feature selection, gene expression data, hybrid approach, microarray.

## I. INTRODUCTION

Cancer was the second leading cause of death worldwide in 2020, causing nearly 10 million deaths. Researchers fear the rates will increase by 50% to 15 million new cases [1], [2]. Cancer starts when abdominal cells grow in organs or tissues of the human body and spreads to its surroundings and, in advanced cases, expands into other organs. Early detection of cancer can help significantly increase survival rates. For the patient to receive appropriate treatment, the kind of cancer must be determined as precisely as possible. The traditional method used microscopic observation on different types of biopsy samples, but this is considered a waste of time and not cost-efficient in advanced cases, and it can produce false negative results. For that reason, the use of DNA microarrays and selection of the correct number of features (genes) is needed to find more predictive and effective genes for cancer classification is essential.

Typically, gene expression data contains a large number of genes, which necessitates the employment of analysis techniques so meaningful information can be obtained [3]. The advent of gene expression technologies has made microarray

data increasingly popular in cancer research classification due to the massive amount of gene expression information (features/genes) that can be used to find common patterns within a set of samples. Microarrays are a prominent method for identifying cancer cells by analyzing the DNA proteins for further analysis of the genes. Microarray data is organized into a matrix called the gene expression matrix, in which each row represents a specific gene and each column indicates an experimental condition [4].

The use of microarray technology can yield useful insights into disease-gene correlations. However, the dimensionality problem, the presence of irrelevant genes, complicates data analysis and cancer classification. To remove unnecessary genes from microarrays and retrieve useful information, a feature selection method and classification algorithm are applied to classify the cancer accurately [5].

Feature selection methods are divided into several categories: filter, wrapper, and embedding. In recent years, a hybrid method has been introduced as part of the general framework for feature selection. The main idea behind feature selection is to choose the most informative and significant genes for the classification problem. This selection can be attained by removing irrelevant genes and noisy data to maximize the correct predictive outcomes for cancer

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang<sup>ID</sup>.

Samples	Genes																			Labels
	0	1	2	...	1998	1999	2000													
0	8589.4160	5468.2407	4263.4077	...	83.5225	28.70125	Tumor													
1	9164.2540	6719.5293	4883.4487	...	44.4725	16.77375	Normal													
...	...	...	...	...	...	...	...													
60	6234.6226	4005.3000	3093.6750	...	32.6875	23.26500	Tumor													
61	7472.0100	3653.9340	2728.2163	...	49.8625	39.63125	Normal													

[62 rows x 2001 columns]

FIGURE 1. Gene expression data matrix.

classification [4]. The hybrid method combines the benefits of both the filter and wrapper techniques. Several hybrid approaches, primarily a merger of filter and wrapper methods or two wrapper methods to identify the useful genes for correct diagnosis, have been developed over the last few years. The hybrid methods integrate the capabilities of both approaches to get the best of both worlds [4].

The goal of this survey is to find contributions to the development of hybrid feature selection methods for cancer classification in recent years.

## II. DNA MICROARRAY GENE EXPRESSION PROFILE

DNA microarrays are a technical alliance of biology and computers that allows for the genome-wide analysis of gene expression in human tissues [3]. DNA microarray technology has been widely used in cancer research for cancer classification. In addition, understanding of the cause of cancer has also made it possible to inspect the expression levels of a large number of genes at the same time [6]. Especially when the technology becomes more widely used and standardized, prices and complexity decrease because of the massive amount of gene expression information (features) that could be used to find common patterns within a set of samples.

The expression level of a gene is represented by the number of gene cells. Gene expression typically yields thousands of genes and a small number of samples. This is an issue in microarrays called high dimensionality. Gene expression also has many useless and superfluous features, and only a few of the evaluated genes may have a significant impact on cancer classification. Genes are coding sections that construct essential building blocks inside the cell and direct proteins to perform a range of functions. The expression variables in the microarray dataset are structured as an  $M \times N$  matrix, where each row contains multiple features each feature is also called a gene, and each column represents samples matrix [4], as shown in Figure 1.

## III. FEATURE (GENE) SELECTION

The main objective of feature selection is to choose the most informative and significant genes for the classification problem. This selection can be attained by removing irrelevant genes that add dimensionality and noisy data to find relevant features and patterns in genes that may help cancer classification. Feature selection offers several advantages [5]:

- Helping researchers visualize, understand, and gain knowledge about the data.



FIGURE 2. Gene filter method flowchart.

- Reducing data and scaling down the storage requirements.
- Generating a simpler model that allows for greater speed and simplicity.
- Improving the performance of the machine learning algorithm.

There are three main feature selection methods used to subset the feature space and help the model perform efficiently: filters, wrappers, and embedded methods. Each method has its own use and way of interacting with the genes. However, two new methods have been added: ensemble and hybrid methods [7]. Many researchers have been applying these new methods to their classification models to generate new feature selection methods. Each of these five methods has distinct characteristics. However, we will explain only those methods most relevant to our project: filter, wrapper, embedded, and hybrid.

### A. FILTER FEATURE SELECTION METHODS

Filter methods, Figure 2, are commonly employed as a pre-processing phase, the earliest step in feature selection to reduce dimensionality. The methods typically calculate a feature/gene relevance score for each feature/gene, rank the features/genes based on their scores, and omit low-scoring features/genes [8]. There are many advantages to using filter methods, the most important being that they achieve more generality with less computational complexity, therefore being suitable for high-dimensional space, and computation is straightforward and fast [8].

The following is an introduction to some of the commonly used filter methods.

- **Information gain (IG)** is a feature selection method based on entropy. It represents how much information is included in a class prediction [9]. Specifically, entropy measures the amount of information in a random variable [10].
- **Mutual information (MI)** measures nonlinear relationships between two random variables by measuring the level of similarity and correlation and then shows how much data can be collected from one random variable by monitoring another random variable X and Y [11]. In other words, MI is a method for identifying features that are highly dependent on all the other features in the same class.
- **Conditional mutual information maximization (CMIM)** is an approach that selects features based on an approximation of that criterion by attempting to reduce the correlation between features/genes [12]. CMIM has the advantage of not selecting features like those already selected, even if they are individually powerful, as it

does not carry additional information about the class. The algorithm starts with empty selected features; it adds features in each iteration, and every new feature is compared using MI with selected features to assess if it is redundant.

- **Minimum redundancy maximum relevance (mRMR)** was proposed by Peng *et al.* in 2003 [13], and it gained popularity in 2019 after Uber became popular [14]. mRMR aims to find the maximum relevance between the features and the target as well as the minimum redundancy between the random variables X and Y. This can be achieved by using the mutual information algorithm [13]. The aim of maximum relevance is to find the most correlated features to the target. The maximum relevance criterion can lead to many redundant features. Therefore, the minimum redundancy routine finds a better subset representation of the whole feature by removing similar features.
- **Random forest ranking (RFR)** is an algorithm that uses decision trees to merge predictions from a collection of random trees [15] by applying accuracy-based ranking. It is based on the correctness of a single tree from a previous random forest evaluation [16].
- **Fast Correlation-Based Filter (FCBF)** is a method for selecting features developed by Yu and Liu [17] for managing multivariate criteria that eliminate noise while maintaining more significant data using symmetrical uncertainty (SU). The FCBF algorithm utilizes several concepts, including predominant correlation and heuristic-heuristic. It identifies a set of features that are highly correlated with the classes and then sorts those values using predominant correlation. A heuristic algorithm is used to remove features that are redundant while keeping those that are more relevant [18].
- **F-score** is also known as Fisher's scoring and the scoring algorithm. It is a selection strategy that takes the F-distribution into account when looking at which individual descriptive features relate to the target features, and based on their scores, each feature is selected independently [19]. The F-score is considered to be simpler than the feature selection algorithm proposed by Chen and Lin [20]. The selection of features subset is based on a small distance between features point from the same target (minimum interclass distance) and a larger distance between different targets (maximum intraclass distance).
- **Relief** algorithm considers the correlation between features, and feature weights are used to select the features to classify. Despite the ease with which the relief technique calculates classification weights, the results can be influenced by noise, which can lead to mistakes in the subset of features acquired [21]. Relief was originally proposed by Kira and Rendell [22].
- **Pareto Optimization** Pareto optimization or multi-objective optimization aim to develop and present a set of acceptable solutions to the decision maker, who



FIGURE 3. Wrapper method flowchart.

will then choose a solution from it. In some cases, a decision-maker can provide additional constraints or criteria either before or after the search to help with guidance, refinement and narrowing, but in this case we will consider the generic scenario in which there is no prior information from the decision-maker [23].

## B. WRAPPER FEATURE SELECTION METHODS

Wrapper methods use a classifier along with learning algorithms to find an optimal subset of features. They have to conduct a search in the space of primary features and select a subset of them. They are known for high computing costs, and they are not suitable for high-dimensional datasets [24]. However, they are more effective than feature-ranking algorithms because they consider the classifier hypothesis [25]. Figure 3 shows the wrapper steps.

Here are descriptions of the most typically used wrapper methods divided by their meta-heuristic categories based on [26]:

### 1) EVOLUTIONARY-BASED: INSPIRED BY EVOLUTIONARY PROCESSES FOUND IN NATURE

- **Genetic Algorithm (GA)** was proposed in 1960 by John Holland. It has been used for many scientific and engineering problems and models, such as optimization, automatic programming, machine learning, economics, immune systems, ecology, population genetics, evolution and learning, and social systems [27]. GA is a heuristic search algorithm inspired by the process of natural evolution and natural selection. The algorithm has three operations: selection, crossover, and mutation. It starts with a selection operation to choose the fittest individuals (genes), discard those that are not well suited for solving the present problem, and pass those chosen to the next generation procedure. This is followed by the crossover operation: new individuals are formed by considering a combination of previously selected individuals. It uses a random selection of two individuals by exchanging the individuals' genes to reduce the number of individuals and select the fittest. Finally, it ends with mutation, which is small random changes to the new solution (individuals) [28].
- **Flower Pollination Algorithm (FPA)** is an algorithm for global optimization based on the population. It follows the pollination behavior of flowering plants and consists of the three main characteristics of pollination: biotic pollination, abiotic pollinations, and flower constancy [29]. The FPA includes a global pollinator and a

local pollinator. The feasible search space is initialized with random vectors after each pollen item is handled as a solution [30].

## 2) SWARM-BASED: INSPIRED ON THE SOCIAL BEHAVIOR OF ANIMALS

- **Artificial Bee Colony Algorithm (ABC)** was proposed in 2005 by Karaboga [31] as a simulation of the foraging behavior of honeybees. There are three types of bees in this algorithm: employed bees, onlookers, and scouts. The number of employed bees is the same as the number of food sources. When the food source found by an employed bee becomes exhausted, this bee becomes a scout. There are three steps repeated in each cycle: (a) the employed bee and the onlookers move to the food sources, (b) the nectar amounts of the food sources are calculated, and (c) the scout bees are recruited and directed to other possible food sources [31].
- **Cuckoo Search (CS)** is an algorithm developed in 2009 by Xin-She Yang and Suash Deb and driven by the cuckoo bird's life cycle. It is based on the behavior of some cuckoo species, such as ani and guira, which engage in obligate brood parasitism by laying their eggs in the nests of other bird species; they may even remove other birds' eggs to increase the chance that their own will hatch [24]. The cuckoo lays one egg at a time and then adds it to a random nest. Then the nest with the highest egg quality is moved to the next generation. CS has a fixed number of nests, and the property that the host bird discovers the egg is

$$p_a \in [0, 1]$$

The bird can then either abandon the nest or get rid of the egg [32].

- **Dragonfly Algorithm (DF)** is inspired by the dynamic and static swarming behavior of dragonflies. The main concept of DF can be understood as a way of estimating the global optimum of an optimization problem [33]. In short, small groups of dragonflies hunt other insects over a small area in a static swarm. The swarming behavior is characterized by local movements and abrupt changes. However, in dynamic swarming, a large number of dragonflies congregate into one swarm and fly for a considerable distance in one direction [34].
- **Moth Flame Algorithm (MFA)** was developed by [35] as a computerized algorithm based on nature. It is primarily inspired by moths' transverse orientation method of navigation in nature. To travel long distances in a straight line, moths maintain a fixed angle with the moon while flying at night. This method works with the moon, which is far away, but it does not work with a close flame.
- **Particle swarm optimization (PSO)** is an algorithm proposed by Kennedy and Eberhart and modeled after the social behavior of bird flocks. It is like birds migrating in flocks toward a common destination, where

intelligence and efficiency come from the cooperation of the flock [36]. PSO uses particles moving in an n-dimensional space to solve an n-variable optimization problem. The particles have fitness values that are evaluated by the fitness function to be optimized and have velocities that control their flight. As the best solutions so far follow the particles, the particles travel through the problem space. PSO starts with a random set of particles—solutions—and then it iterates through the problem space and searches for optimum solutions by updating each generation [37]. PSO is considered one of the better feature selection algorithms, since it can search huge areas cost-effectively in terms of computation. Moreover, it is easier to build and requires fewer parameters [36].

- **Firefly Algorithm (FA)** uses swarm intelligence and upgrades based on a metaheuristic search [38]. Its major strength is solving complex optimization problems. Using FA, the behavior of real fireflies—which is based on the attraction between fireflies, which in turn depends on their brightness—can be simulated. A firefly algorithm must follow the three laws of firefly behavior in a real space [39].
- **Bat Algorithm (BA)** is based on the echolocation behavior of microbats [40]. By utilizing echolocation, microbats can find their prey and distinguish different types of insects in the dark. Bats use short, powerful sound waves to hunt at night and listen for the echo reflecting from a barrier or prey. A bat's particular hearing apparatus can help it determine the size and location of an object [41].
- **Ant Colony Optimization (ACO)** is a heuristic algorithm inspired by the way ants cooperate to find food sources. Each agent in the ACO simulates the real-world behavior of ants as they move from the nest to the food source [42]. The ants move in random directions, depositing a chemical called a pheromone on the ground. When the ants arrive at a path junction, the decision about which path to follow depends on the amount of pheromone on the path. If it is a new path, the probability of the pheromone is the same. However, if the ants have previously chosen one of the crossing paths, the probability that the new ants will follow that path increases. The intensity of the pheromone decreases over time (evaporation), while the amount of pheromone increases with each ant that passes along the path (amplification) [43].

## 3) HUMAN-BASED: TAKES INSPIRATION FROM HUMAN BEHAVIOR AND ACTIVITIES

- **Teaching-Learning-Based Algorithm (TLBO)** was first proposed by Rao *et al.* in 2011 and 2012. The algorithm has two basic modes of learning. In the first, called the teacher phase, the student learns through the teacher, and in the second, called the learner phase, the student learns through interaction with other learners.

The student with the highest grade in the population is chosen to be the teacher during the teacher phase. The teacher is in charge of teaching the learners and raising the class's average grade. During the learner phase, each student is allowed to share their knowledge with other learners randomly to improve their own knowledge. If the other learners have more knowledge than the student, the student will pick up new information; if the other learners do not have more knowledge, the student will not pick up further information. In this stage, the ultimate aim is to raise the class's mean grade. The algorithm can tackle multidimensional, linear, and nonlinear problems with a high degree of efficiency by simulating the teaching-learning process in a classroom: every feature/gene strives to learn from other features/genes to enhance itself [44].

- **Learning Automata (LA)** algorithm was originally designed as an imitation of the learning behavior of biological tissues that can acquire the erratic behavior of their surroundings by frequently interacting with them, thus optimizing the long-term benefits. Action, feedback, learning automata, and a random environment are the four components of the LA learning framework [45].

#### 4) PHYSICS-BASED: INSPIRED BY NATURE'S PHYSICAL PROCESSES

- **Black Hole Algorithm (BHA)** was introduced by Abdolreza Hatamlou in 2013. BHA is population based, inspired by the behavior of black holes, which attract everything around them. The BHA is based on the concept of a black hole, which is a region of space with so much mass concentrated in it that no neighboring object can escape its gravitational pull. Light objects, like everything else that falls into a black hole, cannot escape any BHA iteration, and the best solution is then selected as the black hole, which then attracts other candidates. Stars will be swallowed by black holes if their fitness crosses the event horizon; after that, the search process will start again with a new potential solution star generated at random and placed in the search space [46]. The BHA begins by selecting a random population of possible solutions, called stars. After the initialization step, the population fitness values are evaluated, and the best candidate is chosen as a black hole. The chosen black hole has the best fitness value, and the remaining solutions will move toward the black hole, depending on their position and a random number. During each iteration, the best candidate is considered to be a black hole, and the remaining are treated as stars. Then all the stars near the black hole are absorbed by the black hole. As the stars are moving, if the star reaches a certain position where the cost is less than the black hole, then the star becomes a black hole, and the iteration starts again.
- **Gravitational Search Algorithm (GSA)** originally comes from the laws of Newtonian mechanics, which are

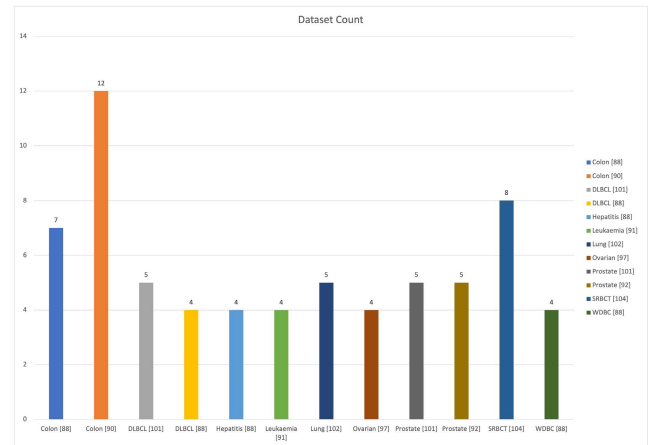


FIGURE 4. Datasets count.

based on an isolated system of masses and their interactions [47]. The GSA takes into consideration gravity and how it attracts other masses [48].

#### 5) MUSIC-BASED: INSPIRED BY MUSIC INSTRUMENT

- **Harmony Search (HS)** In 2001, Zong Woo Geem *et al.* developed the HS [49]. Metaheuristic optimization algorithm based on music. Music is the pursuit of a perfect state of harmony; hence, it was inspired by this observation. The concept of finding harmony in music is analogous to optimizing a process.

#### 6) NO INSPIRATION

- **Crossover** operation involves the mimicking of properties. A random position (crossover point) is selected that separates the parents into two groups. Two new offspring are produced when the parents of the two portions are swapped. This is known as a crossover operation to develop new best options [50].
- **Stacked Autoencoder (SA)** is a deep neural network in which three autoencoder layers (input, output, and hidden) are layered together to form an unsupervised pretraining stage in which an autoencoder's encoder layer is used as the input to the following autoencoder layer [51]. There are two parts to autoencoder training: encoders and decoders. Encoders convert input data into hidden representations, and decoders reconstruct input data from hidden representations [52].

## IV. CLASSIFICATION

Classification is used to determine which dataset the input data originated from. As its name implies, classification in machine learning divides data into multiple categories [53]. The performance of the various algorithms is compared with their results in classification predictive modeling. Classification accuracy is an important metric for assessing how well any model performs based on various predicted classes [54].

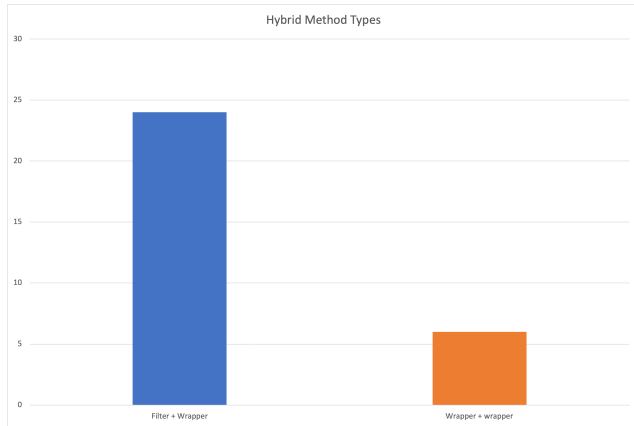


FIGURE 5. Hybrid feature selection method types.

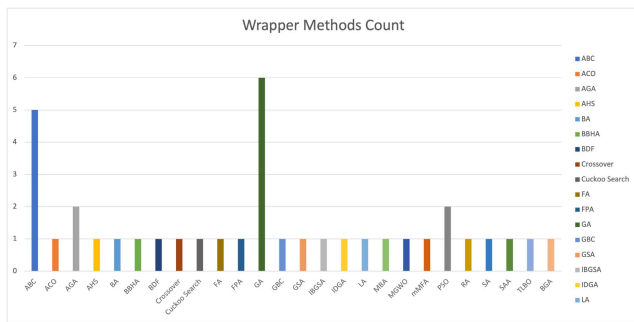


FIGURE 6. Wrapper feature selection methods count.

Below is an overview of the most common classification models used so far.

- **Random Forest (RF)** is an ensemble classification model composed of a set of closely connected decision trees. RF trees are constructed through bagging and random variable selection. Its construction principle is identical to that of decision trees, which is based on recursive partitioning. Each decision tree votes for a class based on its own criteria and variable set, and the classification with the most votes is considered the consensus [55].
- **Support Vector Machine (SVM)** classifies data by identifying a linear or nonlinear separating surface in the input space. A set of support vectors is separated into surfaces that depend only on a subset of the original data. In a high-dimensional space, the SVM constructs a hyperplane or set of hyperplanes that can be used for classification. By using the hyperplane with the greatest distance to the nearest training data points of any class, known as the *functional margin*, a good separation can be achieved. When this functional margin is large, the generalization error of the classifier is small. SVM models are based on a kernel function that transforms the input data into an n-dimensional space in which a hyperplane can be constructed to partition the data [56]. For classification, support vector classifiers are used, while support vector regressions are used when regression data is analyzed [55], [57].

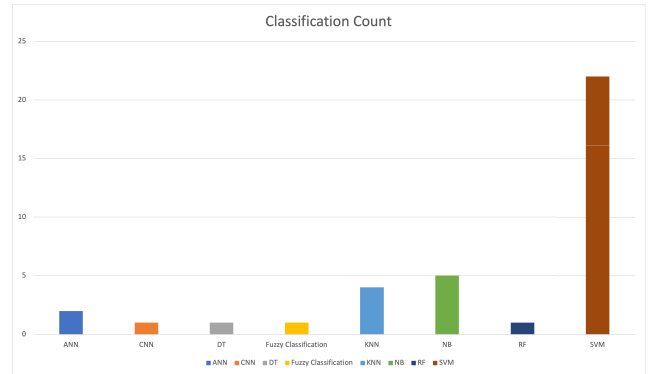


FIGURE 7. Classification methods count.

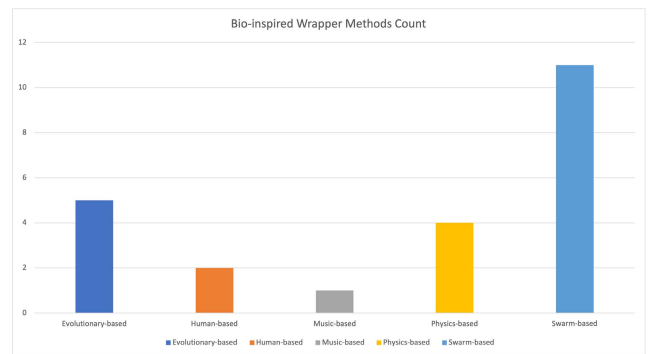


FIGURE 8. Wrapper methods meta-heuristic categories count.

- **Genetic Programming (GP)** is an evolutionary technique for creating computer programs that represent approximate or exact solutions to a problem. GP is merely a subset of GA, with the key difference being the structures of the individuals. Individuals in GA are string structured, those in GP tree structured [58]. GP is based on the evolution of a particular population. In this population, each individual represents a solution to the problem being solved. GP seeks the best solution using a process based on the theory of evolution, where in an initial population of random individuals, after successive generations, new individuals emerge from old individuals through crossover, selection, and mutation. Strong individuals have a better chance of survival to become part of the next generation due to natural selection. Thus, after several generations, the best individual is determined, which corresponds to the final solution of the problem [59].
- **K-Nearest Neighbors (KNN)** is a nonparametric, non-linear, and relatively simple classifier [60]. It classifies a new sample by measuring the “distance” to a set of samples held in memory. The class that the KNN classifier determines for this new sample is determined by the pattern that is most like it (i.e., that has the smallest distance to it). The distance function commonly used in the KNN classifier is the Euclidean distance. A majority voting among the K nearest neighbors is usually performed to select the nearest sample. The parameter K in KNN must be chosen before the classifier is run [59].

- **Naïve Bayes (NB)** is a probabilistic algorithm that employs the Naïve Bayes theorem [61]. In probability theory, the Bayes theorem relates the conditional and marginal probabilities of two random events. It is used to calculate the posterior probabilities of given observations. A Naïve Bayes classifier assumes that features are conditionally independent with respect to class, meaning that the value of a given feature of a class is unrelated to the value of another feature.
- **Artificial Neural Networks (ANN)** McCulloch and Pitts developed it in 1943 [62] its mimics the interaction between nerve cells in the brain by using mathematical and computational techniques, by using this technology translate inputs and outputs to simulate real-world scenarios.
- **Fuzzy classification** is a rules-based classifier that offers substantial benefits concerning functionality, analysis, and design. It involves finding one of such class labels in a set of class labels corresponding to the vector of features of an object. A fuzzy classifier has the advantage of being able to interpret classification rules better than traditional classifiers based on other principles. Its classification accuracy is widely used as a metric of efficiency [63].

## V. HYBRID FEATURE SELECTION METHODS

Hybrid feature selection methods typically combine sequentially and successively two or more feature selection algorithms from different search strategies. It aims to take advantage of both filtering and wrapping techniques to overcome the disadvantage of the individual techniques and reduce the complexity of selecting relevant features from the dataset by reducing the selection time [64].

Hybrid methods developed since 2017 include the following:

- **Intelligent dynamic genetic algorithm (IDGA)** [65]. Developed by Dashtban and Balafar, this is built on the concepts of genetic algorithms, artificial intelligence, random restart hill-climbing, and reinforcement learning. It comprises two steps. First, the dataset is filtered using the Fisher score method to choose the top N statistically significant genes for the next step, and two alternative scoring techniques, the Fisher score and the Laplacian score, are applied. Second, the IDGA method is then used to examine the significant gene subset using an SVM classifier [66]. In addition, it provides the required crossover and mutation probability as well as faster convergence for the recognition of predictive genes [67].
- **Genetic Bee Colony Algorithm (GBC)** [68]. Alshamlan *et al.* proposed a new hybrid meta heuristic feature selection. It was built using the ABC and GA algorithms, which were both bio-inspired. The goal is to choose the genes that are most significant in attempt to optimise the classifier's accuracy. The authors of GBC combine GA operators with the ABC algorithm to produce a controlled optimization approach based on the modified ABC algorithm.
- **RFR-IDGA-RF** [66]. Proposed by Pashaei and Pashaei, RFR-IDGA-RF is a new hybrid approach that employs both random forest ranking (RFR) as a filter method and the intelligent dynamic genetic algorithm (IDGA). RFR is used to pick only the important variables (genes) and their accommodating high-dimensional genomic data to eliminate unwanted genes from a new subset and its fitness function. IDGA is used to find the most informative subset from the produced subset in the filter method. The RF classifier, with leave-one-out cross-validation (LOOCV), is used in both fitness and classification of the final top genes, since it has higher performance than the SVM classifier for microarray classification, which is used to evaluate two cancer types of datasets (colon and leukemia) to select the most meaningful genes. The Fisher score ranking method is used to compare the results since the number of genes needed to reach higher accuracy is ambiguous. The Fisher score for the leukemia dataset was not significantly different, but for the colon dataset, it was significantly different. The end experimental results have shown a 100% accuracy rate for leukemia and 95.16% for colon cancer; the authors argue that based on recently published work, their model is highly accurate and selects fewer genes.
- **mRMR-BBHA** [69]. Proposed by Pashaei and Pashaei, mRMR-BBHA is a hybrid method that combines minimum redundancy maximum relevance (mRMR) with the binary black hole optimization algorithm (BBHA) to filter out noisy data and select highly discriminative genes. It was also used with the SVM classification model to accurately diagnose cancer genes. mRMR was used to find the most suitable attributes based on their relevance to the class tags, and at the same time, it minimizes the repetition between attributes. BBHA was also employed as a search algorithm that mimics the behavior of a black hole. It has been applied to two benchmark cancer datasets, colon and breast, showing higher accuracy than SVM with mRMR alone while using a small number of gene subsets. For example, the breast dataset has selected an average of 22.5 genes to achieve 94.48% accuracy, while the colon dataset has achieved a classification accuracy of 98.87% with an average of only 10.33 selected genes.
- **MIMAGA-Selection** [70]. Lu *et al.* proposed MIMAGA-Selection, a new hybrid feature selection algorithm, by merging mutual information maximization (MIM), which identifies genes that are heavily reliant on other genes in the same category, with the adaptive genetic algorithm (AGA) to obtain the highest possible level of optimal results by determining the most appropriate crossover probability and mutation probability values. MIMAGA-Selection's primary

**TABLE 1.** Comparison of the performance of recent and relevant proposed algorithms.

Reference	Hybrid Selection Methods		Classification Method	Dataset	Accuracy	No. of selected genes
	Filter Method	Wrapper Method				
[68]	Mutual Information Maximization (MIM)	Adaptive Genetic Algorithm (AGA)	Support Vector Machine (SVM)	Colon [91]	83.41%	202
				Leukemia [94]	97.14%	198
				Prostate [95]	97.31%	205
				Lung [96]	94.67%	216
				Breast [97]	95.21%	216
				SRBCT [97]	88.64%	207
[69]	Recursive Feature Elimination (RFE)	Binary Dragonfly (BDF)	Support Vector Machine (SVM)	Breast [98]	86.22%	7237
				Colon [91]	97.46%	510
				DLBCL[99]	89.44%	1210
				Leukemia [94]	95.81%	1522
				Lung [96]	99.14%	3737
				Ovarian [100]	98.19%	4573
[70]	Mutual Information (MI)	Genetic Algorithm (GA)	Support Vector Machine (SVM)	Colon [91]	100%	20
				Lung [96]	99.21%	20
				Ovarian [100]	80.39%	20
[17]	Fast Correlation based Filter (FCBF)	Particle Swarm Optimization (PSO)	Support Vector Machine (SVM)	Hepatitis [92]	86.11%	8
				Arrythmia [92]	78.08%	134
				Colon [91]	96.3%	999
				DLBCL [92]	100%	3204
				WDBC [92]	98.82%	15



**TABLE 1. (Continued.) Comparison of the performance of recent and relevant proposed algorithms.**

Reference	Hybrid Selection Methods		Classification Method	Dataset	Accuracy	No. of selected genes
	Filter Method	Wrapper Method				
[71]	-	Learning Automata (LA)	Support Vector Machine (SVM)	Colon [91]	100%	8
				Colon [91]	100%	9
				Colon [91]	100%	10
				ALL_AML [101]	100%	2
				SRBCT [101]	99.8%	4
				SRBCT [101]	99.38%	5
		Genetic Algorithm (GA)		SRBCT [101]	99.94%	6
				MLL [101]	95.71%	3
				Tumors_9 [102]	85.42%	8
				Tumors_9 [102]	89.15%	10
				Tumors_11 [102]	84.65%	8
				Tumors_11 [102]	85.23%	10
[79]	Minimum Redundancy Maximum Relevancy (MRMR)	Flower Pollination Algorithm (FPA)	Support Vector Machine (SVM)	Colon [91]	88.01%	3.56
				Ovarian [92]	100%	4
				Breast [92]	85.88%	16.8
[78]	-	Genetic Algorithm (GA)	Support Vector Machine (SVM)	WDBC [92]	99.41%	13
				Colon [91]	81.67%	1027
		Artificial Bee Colony (ABC)		Hepatitis [92]	91.67%	9
				DLBCL [92]	91.45%	1943

TABLE 1. (Continued.) Comparison of the performance of recent and relevant proposed algorithms.

Reference	Hybrid Selection Methods		Classification Method	Dataset	Accuracy	No. of selected genes
	Filter Method	Wrapper Method				
[17]	Fast Correlation based Filter (FCBF)	Genetic Algorithm (GA)	Support Vector Machine (SVM)	Hepatitis [92]	81.94%	16
				Arrhythmia [92]	85.90%	245
				Colon [91]	96.3%	1536
				DLBCL [92]	100%	2330
				WDBC [92]	99.02%	25
[72]	Ensemble Gene Selection (EGS) + F-score	Adaptive Genetic Algorithm (AGA)	Support Vector Machine (SVM)	Colon [103]	98.9%	16
				SRBCT [93]	98.79%	13
				Breast [103]	98.79%	17
				Lung [93]	99.03%	14
				DLBCL [93]	99.01%	18
				Leukemia [103]	98.72%	13
[71]	-	Genetic Algorithm (GA)	Support Vector Machine (SVM)	Leukemia [104]	77.36%	NAN
				Brain-tumor1 [104]	76.9%	NAN
				Brain-tumor2 [104]	73.79%	NAN
		Artificial Bee Colony Algorithm (ABC)		Lung [104]	96.3%	NAN
				Prostate [104]	77.2%	NAN
				DLBCL [104]	98.91%	NAN

**TABLE 1.** (Continued.) Comparison of the performance of recent and relevant proposed algorithms.

Reference	Hybrid Selection Methods		Classification Method	Dataset	Accuracy	No. of selected genes
	Filter Method	Wrapper Method				
[73]	Robust Minimum Redundancy Maximum Relevancy (rMRMR)	Modified Bat Algorithm (MBA)	Support Vector Machine (SVM)	Breast [104]	95.4%	12.63
				MLL [104]	100%	8
				Colon [104]	97.85%	12.27
				ALL_AML [104]	100%	4.07
				ALL_AML_3c [104]	100%	5.33
				ALL_AML_4c [104]	100%	6.73
				Lymphoma [104]	100%	8.13
				CNS [104]	100%	11.2
				Ovarian [104]	100%	3.07
				SRBCT [104]	100%	9.13
[37]	-	Crossover	K-Nearest Neighbors (KNN)	Leukemia [105]	97.98%	NAN
				Prostate [105]	99.51%	NAN
				Colon [105]	98.54%	NAN
		Cuckoo Search		Lung [105]	99.82%	NAN
				Lymphoma [105]	99.96%	NAN
[64]	Random Forest Ranking (RFR)	Intelligent dynamic genetic algorithm (IDGA)	Random Forest (RF)	Colon [91]	95.16%	4
				Leukemia [92]	100%	7

TABLE 1. (Continued.) Comparison of the performance of recent and relevant proposed algorithms.

Reference	Hybrid Selection Methods		Classification Method	Dataset	Accuracy	No. of selected genes
	Filter Method	Wrapper Method				
[75]	F-score	Firefly Algorithm (FA)	Support Vector Machine (SVM)	Colon [91]	94.3%	15
				Lung [106]	100%	2
				Leukemia_1 [94]	100%	5
				SRBCT [107]	100%	8
				Leukemia_2 [108]	97.8%	10
[83]	Mutual Information Maximization (MIM)	Modified Moth Flame Algorithm (mMFA)	Support Vector Machine (SVM)	Leukaemia1 [94]	100%	6
				DLBCL[99]	100%	11
				Prostate [95]	100%	14
				CNS [111]	100%	13
				Colon [91]	100%	20
				Breast [111]	91.75%	11
				Ovarian [100]	98.42%	26
[84]	Mutual Information (MI)	Improved Binary Gravitational Search Algorithm (IBGSA)	Support Vector Machine (SVM)	ALL-AML [104]	99.71%	NAN
				Colon [104]	90.90%	NAN
				Ovarian [104]	99.64%	NAN
				GSE4115 [112]	72.92%	NAN
				GSE10245 [112]	98.33%	NAN
[82]	Mutual Information (MI)	Ant Colony Optimization (ACO)	Fuzzy Classification	Colon [110]	100%	NAN
				Leukaemia [110]	100%	NAN
				Prostate [110]	90.85%	NAN

**TABLE 1.** (Continued.) Comparison of the performance of recent and relevant proposed algorithms.

Reference	Hybrid Selection Methods		Classification Method	Dataset	Accuracy	No. of selected genes
	Filter Method	Wrapper Method				
[76]	Conditional Mutual Information Maximization (CMIM)	Binary Genetic Algorithm (BGA)	Support Vector Machine (SVM)	Colon [91]	66.23%	21
				SRBCT [107]	82.74%	21
				Lung [105]	72.58%	28
				DLBCL [104]	79.74%	25
				Prostate [104]	91.54%	19
			K-Nearest Neighbors (KNN)	Colon [91]	82.19%	21
				SRBCT [107]	61.24%	21
				Lung [105]	61.27%	28
				DLBCL [104]	83.24%	25
				Prostate [104]	89.08%	19
			Decision Tree (DT)	Colon [91]	83.13%	21
				SRBCT [107]	82.91%	21
				Lung [105]	84.04%	28
				DLBCL [104]	89.91%	25
				Prostate [104]	92.75%	19
			Naive Bayes (NB)	Colon [91]	78.42%	21
				SRBCT [107]	71.23%	21
				Lung [105]	94.63%	28
				DLBCL [104]	74.23%	25
				Prostate [104]	81.43%	19

TABLE 1. (Continued.) Comparison of the performance of recent and relevant proposed algorithms.

Reference	Hybrid Selection Methods		Classification Method	Dataset	Accuracy	No. of selected genes
	Filter Method	Wrapper Method				
[77]	-	Gravitational Search Algorithm (GSA)	Naive Bayes (NB)	Leukaemia_2 [93]	98.84%	12
				Colon [93]	98.87%	16
				DLBCL [93]	99.62%	17
				SRBCT [93]	99.17%	11
				Leukaemia_1 [93]	94.15%	16
		Teaching Learning-based Algorithm (TLBO)		Lung [93]	99.61%	13
				Brain tumor_1 [93]	96.92%	15
				11_tumor [93]	93.04%	13
				9_tumor [93]	70.88%	12
				Prostate [93]	98.42%	7
[78]	-	Particle Swarm Optimization (PSO)	Support Vector Machine (SVM)	WDBC [92]	96.71%	16
		Artificial Bee Colony (ABC)		Colon [91]	79.17%	599
				Hepatitis [92]	87.5%	4
				DLBCL [92]	90%	1383
[79]	Minimum Redundancy Maximum Relevancy (MRMR)	Genetic Algorithm (GA)	Support Vector Machine (SVM)	Colon [91]	89.41%	6.73
				Ovarian [92]	100%	5.87
				Breast [92]	85.88%	22.23
[80]	Relief	Stacked Autoencoder Approaches	Convolutional Neural Networks (CNN)	Ovarian [100]	98.6%	NAN
				Leukaemia [94]	99.86%	NAN
				CNS [109]	83.95%	NAN
			Support Vector Machine (SVM)	Leukaemia [94]	97.1%	3
				Prostate [95]	94.1%	6
				SRBCT [107]	85%	6

**TABLE 1. (Continued.) Comparison of the performance of recent and relevant proposed algorithms.**

Reference	Hybrid Selection Methods		Classification Method	Dataset	Accuracy	No. of selected genes
	Filter Method	Wrapper Method				
[81]	Fisher Criterion	Bat Algorithm (BA)	Support Vector Machine (SVM)	Leukaemia [94]	97.1%	3
				Prostate [95]	94.1%	6
				SRBCT [107]	85%	6
			K-Nearest Neighbors (KNN)	Leukaemia [94]	100%	3
				Prostate [95]	97.1%	6
				SRBCT [107]	100%	6
			Naive Bayes (NB)	Leukaemia [94]	100%	3
				Prostate [95]	97.1%	6
				SRBCT [107]	100%	6
[85]	Independent component analysis (ICA)	Artificial Bee Colony (ABC)	Naive Bayes (NB)	Colon [91]	98.17%	16
				Leukaemia [94]	98.18%	12
				Prostate [95]	98.38%	16
				High-grade glioma [112]	94.39%	9
				Lung [93]	92.76%	24
				Leukemia 2 [98]	97.12%	15
[86]	robust Minimum Redundancy Maximum Relevancy (rMRMR)	Modified Gray Wolf Optimizer (MGWO)	Support Vector Machine (SVM)	Colon [91]	95.86%	9.8
				Lung [93]	97.91%	15.8
				ALL-AML-3C [104]	100%	6.7
				ALL-AML-4C [104]	99.9%	11.36
				SRBCT [107]	100%	12.3
				CNS [109]	99.38%	17.46
				ALL-AML [104]	100%	5.06
				MLL [101]	100%	8.4
[67]	Minimum Redundancy Maximum Relevance (mRMR)	Binary Black Hole Optimization Algorithm (BBHA)	Support Vector Machine (SVM)	Breast [93]	95.70%	21

**TABLE 1. (Continued.) Comparison of the performance of recent and relevant proposed algorithms.**

Reference	Hybrid Selection Methods		Classification Method	Dataset	Accuracy	No. of selected genes
	Filter Method	Wrapper Method				
[87]	Pareto Optimization	Adaptive Harmony Search (AHS)	Support Vector Machine (SVM)	Leukemia [94]	100%	3
				Prostate [95]	100%	4
				DLBCL[99]	100%	6
				Colon [91]	81%	5
			Naive Bayes (NB)	Leukemia [94]	100%	3
				Prostate [95]	98%	4
				DLBCL[99]	88%	6
				Colon [91]	80%	5
			K-Nearest Neighbors (KNN)	Leukemia [94]	99%	3
				Prostate [95]	93%	4
				DLBCL[99]	96%	6
				Colon [91]	78%	5
[89]	Minimum Redundancy Maximum Relevancy (MRMR)	Simulated Annealing (SA)	Support Vector Machine (SVM)	Ovarian [100]	99.15%	6
				Colon [91]	97.02%	9
				Leukemia [94]	97.65%	7
		Rao Algorithm (RA)		Lung [113]	90.22%	5
				Lymphoma-3 [111]	98.44%	6
				ALL-AML-3 [111]	98.02%	7
				SRBCT [111]	99.81%	5
[88]	Independent component analysis (ICA)	Artificial Bee Colony (ABC)	Artificial Neural Networks (ANN)	Colon [91]	97.34%	12
				Leukaemia [94]	98.21%	12
				Prostate [95]	97.88%	20
				High-grade glioma [112]	93.22%	9
				Lung [93]	94.78%	15
[88]	Independent component analysis (ICA)	Genetic bee colony (GBC)	Artificial Neural Networks (ANN)	Colon [91]	98.22%	9
				Leukaemia [94]	99.03%	8
				Prostate [95]	98.45%	12
				High-grade glioma [112]	96.43%	9
				Lung [93]	97.22%	10



purpose is to minimize the dimension of the gene expression profile and eliminate duplicated genes. The proposed algorithm was tested against six benchmark gene datasets: leukemia, colon, prostate, lung, breast, and small-round-blue-cell tumor (SRBCT). The SVM was selected as a classifier, and 30 repetitions of the classification process were performed. On the same dataset with the same target gene number, the authors used three existing algorithms—sequential forward selection (SFS), ReliefF, and MIM—with the SVM classifier to compare the accuracy of the MIMAGA-Selection algorithm. MIMAGA-Selection's accuracy was higher than that of existing feature selection algorithms, according to the results. Moreover, the authors used four different classifiers to classify the selected gene using the MIMAGA-Selection algorithm. The accuracy of all four classifiers was greater than 80%.

- **Hybrid SVM-RFE and BDF [71].** Medjaheda *et al.* proposed a hybrid of SVM-RFE and binary dragonfly (BDF). It was used to diagnose cancer using a BDF algorithm with the SVM. Because the efficiency of SVM-RFE alone differs based on the genes, and redundancies in genes are not considered, BDF was applied as a wrapper method to solve those issues. SVM-RFE was used to select the most ideal gene from the datasets and eliminate the remaining nonapplicable features. SVM-REF eliminated 40% of the features, and the remaining subset was processed via binary dragonfly (BDF) to enhance the performance by obtaining only the informative genes. The proposed algorithm was applied to six microarray cancer datasets and achieved comparable results, but for breast cancer, it was not adequate since it achieved high accuracy but with a very large number of features.
- **Two-stage MI-GA [72].** Rani and Devaraj designed a two-stage MI-GA feature selection technique for cancer classification aimed at extracting important gene mutation groups suitable for the cancer classification of patients with ovarian, lung, and colorectal cancer. The technique uses a pipeline of MI and a GA for gene selection from a gene expression dataset. In the first stage, MI-based gene selection is applied to select only the genes with high information related to the class. In the second stage, the resultant feature subset from the first stage is used to identify and select the final optimal set by applying the GA. The feature subset from the two-stage method was evaluated using SVM-based classification on different types of cancer (colon, lung, and ovarian) datasets, where the highest classification accuracy, 96.77%, appeared in the colon cancer dataset, with only 10 extracted genes.
- **GALA [73].** Motieghader *et al.* proposed a mixed cancer classification hybrid algorithm that uses the GA as one wrapper combined with the LA as another wrapper. The GA was first used to assign a score for each chromosome by the gene locations. After that, the chromosome that had the best or highest score with the maximum fitness function value was located. GALA was applied to the SVM classifier to predict cancer. The authors chose the SVM classifier as the classification model. The proposed approach was applied to six different binary and multiclass microarray cancer datasets—colon, ALL\_AML, SRBCT, MLL, tumors\_9, and tumors\_11—and performed well. Its mean classification accuracy on the colon dataset with 8 genes was 99.46%; the ALL\_AML, SRBCT, and MLL datasets had mean classification accuracies of 100%, 97.35%, and 93.96% when selecting 2, 4, and 3 genes, respectively, and the tumors\_9 and tumors\_11 datasets' mean classification accuracies with 10 genes were 86.52% and 84.38%, respectively.
- **FCBF-GA and FCBF-PSO [18].** Djellali *et al.* proposed two new hybridized filter and meta-heuristic methods for optimal feature selection. The first one combines the fast correlation-based filter (FCBF), known to be powerful in removing unneeded and irrelevant features, with the GA. This hybrid method has two steps. The first method, FCBF-GA, uses the FCBF to eliminate unneeded features, and the GA selects features that the FCBF has already selected with other features since two features may be compatible and yield greater accuracy when used together. The second method, FCBF-PSO, combines the FCBF, which reduces features that are not necessary or useful, with particle swarm optimization (PSO), whose global optimization ability is well known in large search areas and whose computational complexity is relatively low. The experiments were conducted on five cancer microarray datasets (Wisconsin Diagnostic Breast Cancer, colon, hepatitis, diffuse large B-cell lymphoma, and lung) using the SVM as the classifier. The results showed that the second method, FCBF-PSO, surpasses FCBF-GA and other existing methods when the accuracy and number of selected genes are considered.
- **Hybrid EGS + F-score with AGA [74].** Shukla *et al.* were driven to develop a new hybrid gene selection strategy that helps decrease false positives and correctly categorize cancers in a short time. The proposed hybrid method, EGS + F-score with AGA, consists of two phases. The first phase utilizes the external guide sequence (EGS) method, which uses a multi-layered approach, and the F-score approach is used to filter noise and redundant genes from the dataset. In the second phase, an adaptive genetic algorithm (AGA) acts as a wrapper and identifies important subsets of the gene from the resulting reduced datasets produced by the EGS to help detect cancer or tumors. The developed model was tested on six cancer gene datasets (colon, breast, diffuse large B-cell lymphoma, SBRCT, lung, and leukemia), and the outcome of the experiment reveals high accuracy, greater than 98%, for all cancer gene datasets.

- **rMRMR-MBA [75]**. Al-Betara, Alomari, and Abu-Romman tried to solve an issue facing many selection methods, finding the most important and dependable genes, and proposed rMRMR-MBA, a hybrid filter and wrapper approach with a filter stage and a wrapper stage. At the filter stage, robust minimum redundancy maximum relevancy (rMRMR) will select the most promising genes by giving scores for each gene. At the wrapper stage, a modified bat algorithm (MBA) will act as a search engine and sort the genes by their scores to identify a small set of informative features. rMRMR-MBA was evaluated on 10 cancer gene expression datasets (breast, MLL, colon, ALLAML, ALLAML-3C, ALLAML-4C, lymphoma, CNS, ovarian, and SRBCT); accuracy was 100% for 8 datasets (MLL, ALLAML, ALLAML-3C, ALLAML-4C, lymphoma, CNS, ovarian, and SRBCT), 97.65% for the colon dataset, and 95.4% for the breast dataset. What makes the proposed hybrid method promising is that the number of selected genes is less than 10 for the 8 datasets that reached 100% accuracy.
- **GAABC [76]**. Ge *et al.* wanted to solve the dimensionality problem of microarray data classification. Their hybrid method, GAABC, merges the artificial bee colony (ABC) algorithm with the GA to enhance the GA's ability to jump from local to global search functionality and increase the diversity of bee populations. The ABC has known defects of premature convergence and early fall into local extremum, which is why it was combined with the GA. The experimental results showed 80% accuracy, lower than other existing proposed hybrid methods.
- **CSC [50]**. Sampathkumar *et al.* proposed a new hybrid bio-inspired algorithm combining two wrapper methods: cuckoo search, which is used to find the significant cancer-causing genes, with a crossover operator, which is a useful technique for luring populations away from local minima that solve the cuckoo search issue. The authors applied the cellulose synthase complex (CSC) selection method to the KNN classifier. Five cancer gene expression datasets (prostate, colon, leukemia, lung, and lymphoma) were used in the experiment. The results have shown that the CSC surpasses other well-known methods, yielding a classification accuracy of 99% for the prostate, lung, and lymphoma datasets, 96.98% for leukemia, and 98.54% for colon.
- **FFF [77]**. Almgren and Alshamlan improved a previously developed wrapper to generate a new hybrid selection method called fuzzy firefly (FFF), which consists of a filtering phase and a gene selection phase. In the filtering phase, the F-score is used to reduce the dimensionality of the data and reduce the complexity of the search area. In the gene selection phase, a wrapper method called FF is applied to locate the genes that are more informative. The experiments were carried out on five microarray cancer datasets (leukemia\_2, SRBCT, lung, leukemia\_1, and colon) having both binary and multi-class labels. Experimental results show that FFF-SVM has 100% accuracy for the lung, leukemia\_1, and SRBCT datasets, in which the number of selected genes was less than 10. The accuracy for the leukemia\_2 dataset was 97.8%, and 94.3% for colon.
- **CMIM + BGA [78]**. Shukla *et al.* proposed CMIM + BGA for addressing the problem of classification, as well as the limitations of present methods. It used conditional mutual information maximization (CMIM) for selecting the feature subset with the highest score to reduce diagnosis time when a large number of noisy, redundant genes are removed and the binary genetic algorithm (BGA) to accelerate the process of locating key feature subsets. The CMIM method can reduce diagnosis time when a large number of noisy, redundant genes are removed. The effectiveness of the hybrid CMIM + BGA algorithm was evaluated using a number of classifiers on five biological datasets and five University of California at Irvine datasets of different dimensionalities and several instances. The authors ran filter-wrapper feature selection on four different classifications (SVM, DT, KNN, and NB). According to the findings of the evaluation, the proposed method provides adequate support for major feature reduction and outperforms existing methods; the classification accuracy of KNN scored was the lowest precision, 40.04%, the SVM the highest, 99.32%; the SRBCT dataset has the lowest classification accuracy, 61.24%, the diffuse large B-cell lymphoma (DLBCL) the best, 99.32%.
- **TLBOGS [79]**. Shukla, Singh, and Vardhan wanted to lessen the dimensionality issue in microarray gene data and increase the interpretability of discriminative gene data, so they produced the TLBOGS hybrid wrapper method that integrates the properties of the teaching learning-based algorithm (TLBO) and the gravitational search algorithm (GSA). The TLBO, introduced in 2011, offers great potential for identifying gene subsets with near-optimal properties in high-dimensional spaces, but it has a few limitations, such as premature and slow convergence. To solve these issues, the GSA is used since it has excellent global search capability. The developed method was tested on ten microarray cancer datasets (leukemia\_2, colon, DLBCL, SRBCT, lung, prostate, brain tumor\_1, 11\_tumor, and 9\_tumor). The results of the experiments show that the proposed method has higher classification accuracy and a more optimal number of feature sets than current approaches. The proposed method achieves greater than 98% accuracy in six datasets (leukemia\_2, colon, DLBCL, SRBCT, lung, prostate), with the greatest accuracy, 99.62%, in the DLBCL dataset.
- **ABC-PSO and ABC-GA [80]**. Djellali *et al.* proposed two hybrid methods based on the artificial bee colony (ABC). The first method, ABC-PSO, combined the ABC with particle swarm optimization (PSO)

to improve the search capability of the ABC bees when they found no food source and to give greater stability between exploration and exploitation. The second method, ABC-GA, combines ABC with the genetic algorithm (GA) to find a high balance between exploration and exploitation since each chromosome—possible solution—and the collection of chromosomes form a population. In the onlooker and scout phases, GA mutation operators are used. The experimental results indicate that the proposed hybrid ABC-GA method is competitive with existing methods and outperforms ABC-PSO in identifying and classifying the Wisconsin Diagnostic Breast Cancer, colon, hepatitis, and DLBCL cancer datasets with the smallest number of features. The results of the experiments illustrate the effectiveness of mutation operators in terms of accuracy and particle swarm for smaller characteristics. Although ABC-PSO has the lowest accuracy, this hybrid method tends to produce fewer gene features.

- **MRMR-FPA and MRMR-GA [81]**. Alomari *et al.* wanted to solve the gene selection issue since the sheer number of genes and the small number of patient samples make it difficult for classifiers to produce appropriate classification results. The majority of these genes are repetitious and unnecessary, which may impair categorization. The authors proposed MRMR-FPA, which consists of minimum redundancy maximum relevancy (MRMR) as the filter method and the flower pollination algorithm (FPA) as the wrapper method to determine the most informative gene subset. To evaluate the MRMR-FPA, the authors developed another method, the MRMR-GA, which is based on MRMR as the filter method and the GA as the wrapper method. The experiments were conducted on three microarray cancer gene datasets (colon, ovarian, and breast). The performance of MRMR-FPA and MRMR-GA were similar on the ovarian and breast datasets. However, the MRMR-GA had a higher classification accuracy on the colon dataset. The comparison of MRMR-FPA with MRMR-GA revealed that MRMR-FPA was able to achieve similar classification accuracy with a lower number of genes selected. This gives MRMR-FPA the potential for overcoming the gene selection issue.
- **Relieff [82]**. Kilicarslan, Adem, and Celik developed a hybrid method called Relieff for dimension reduction and classification that combines the Relief method with the stacked autoencoder. Relief ensures that data is compressed to save storage space, and it reduces computing complexity, but this method often results in data loss; this loss can be solved by applying the stacked autoencoder as a wrapper to acquire new characteristics from the outputs of the hidden layers. Relieff was then used with convolutional neural networks (CNNs) for classification. The developed method was tested on the ovarian, leukemia, and CNS microarray datasets, in which it had classification accuracies of 98.6%, 99.86%, and 83.95%, respectively.
- **MOBBA-LS [83]**. The authors proposed MOBBA-LS, a novel bio-inspired multiobjective algorithm that aims to identify the informative genes that employ the bio-inspired multi-objective binary bat algorithm (MOBBA) by using specific local searches based on the BA with a Fisher criterion that aims at identifying the informative genes. MOBBA-LS uses the fast-non-dominated-sort algorithm to locate the leader bats, which are the ultimate solution and are theoretically participants in the first front of the multi-objective outcome. The proposed method was tested against three different microarray cancer datasets: leukemia, SRBCT, and prostate. The proposed method achieved the best accuracy in the prostate cancer dataset while using a much smaller number of genes.
- **Hybrid stem cell (HSC) [84]**. For constructing fuzzy classification systems, Vijay and Ganeshkumar developed a novel hybrid stem cell (HSC) algorithm that combines ant colony optimization (ACO) and the stem cell algorithm with MI, which is a strategy for extracting the most informative genes from a large microarray dataset. MI is used first to reduce the gene dimensionality. Using the ACO algorithm, the HSC rule set is represented by integer values. The simulated performance results of the proposed approach were validated using several microarray datasets. These findings show that the proposed HSC algorithm generates a more precise fuzzy system than existing methods.
- **MIM-mMFA [85]**. Dabba *et al.* combined the modified moth flame algorithm (mMFA) and mutual information maximization (MIM) to build the MIM-mMFA to solve gene selection in microarray data classification. As a prefilter, MIM is used to determine the significance of the genes and remove duplicated genes, and the mMFA is used to select gene subsets and score them based on fitness scores determined by an SVM with LOOCV.
- **MI-IBGSA [86]**. Yan *et al.* used MI to rank and select features for the wrapper method's population based on their significance. The gravitational search algorithm (GSA) is then employed to find an optimal feature subset based on its efficiency. While the GSA has limitations in terms of its search speed and premature convergence, it remains a powerful optimization algorithm.
- **ICA + ABC + NB [87]**. Musheer *et al.* developed a novel feature selection methodology, which consists of two steps: the Independent component analysis (ICA) extraction method and the Artificial Bee Colony (ABC) wrapper approach, with Naive Bayes (NB) as a classifier. A major advantage of ICA is that the number of extracted features is always equal to the number of samples in the dataset. ICA has this issue that it do not know which subset is the best subset of features. To solve this issue the authors used ABC as a wrapper method to select the best subset of features.

- **rMRMR-MGWO [88]**. A new gene selection for microarray data classification has been developed called rMRMR-MGWO in which its compose of two phases, filter and wrapper methods, as a filter, Robust Redundancy Minimum Maximum Relevancy (rMRMR) was applied and Modified Gray Wolf Optimizer MGWO) with SVM classifier was used as a wrapper to find the optimal subset of genes. the authors improved the GWO with TRIZ optimization mechanism to improve the exploration ability of the wolves to select the important genes to increase the classifier classification.
- **Pareto Optimization + AHS [89]**. As a solution to the high-dimensional dataset issue, Dash proposed a two-stage hybrid feature selection method. An AHS-based probability distribution factor was used to determine the optimal gene ranking in the first stage. To select a minimum number of top-ranked genes, Pareto Optimization was applied as a feature selection method during the second stage. To evaluate the proposed method and check which classifier gives the best results, three classifiers (KNN, NB and SVM) were used. Results show that the SVM classifier provides better results than other classifiers, which give 100% accuracy to most datasets.
- **ICA + ABC + ANN and ICA + GBC + ANN [90]**. In this paper, the authors examine artificial neural networks (ANNs) with two different hybrid algorithms. It combines Independent component analysis (ICA), an algorithm used commonly in filtering, with two bio-inspired approaches: Artificial Bee Colony (ABC) and Genetic Bee Colony (GBC). Dataset dimensionality was reduced using ICA. Five different datasets were analyzed to test the proposed algorithm's performance. According to the findings section, ICA + GBC produced higher accuracy and select lower genes number for the microarray datasets.
- **mRMR-SARA [91]**. Santos Kumar Baliarsingh combines Simulated Annealing (SA) and the Rao Algorithm (RA). In the hybrid technique, SA handle local search and RA handle global optimization. To select relevant gene subsets from the microarray dataset, the proposed method uses an algorithm known as minimum redundancy maximum relevance (mRMR). This method was tested on five binary-class and multiclass datasets. The authors found out that due to the addition of RA optimization to the training method, the accuracy of the model increased.

## VI. ANALYSIS AND DISCUSSION

The purpose of this overview of previous studies is to investigate the current research in hybrid gene feature selection approaches and learn about the current research community's tendencies for the five-year period up to 2020.

Based on the recent studies table, and for comparing the results of previous studies, we must keep in mind the studies with similar datasets, feature selection, and classifiers.

Moreover, it seems that the previous studies show fewer similarities. However, we conclude the following:

- 1) There are many cancer datasets used to predict using genes whether a person could develop cancer, and the most-used cancer dataset types are, in order, colon [90], SRBCT [88], colon [88], as shown in Figure 4.
- 2) Figure 5 shows that most of the studies listed combined both filter and wrapper feature selection methods into new hybrid methods. The filter method was used to improve the performance of the wrapper method, except in four studies, which used only a wrapper method.
- 3) Two studies, [72] and [73], achieved 100% accuracy with the colon [92] dataset with 20, 8, 9, and 10 genes. However, when the proposed method was tested with other datasets, the accuracy was less than 100%. Also, when the same colon [92] dataset was used in other studies, the accuracy was 83% in [70], 97% in [71], 94% in [77], and 60%–83% in [78]. This might be due to the large number of selecting genes as [70] and [71] select 202 and 510 genes, respectively, while [77] selects 15 and [78] selects 21; they achieved lower accuracy, and that might contribute to the chosen feature selection methods, since all studies that have used colon [92] use the SVM classifier.
- 4) Among all other wrapping methods, the genetic algorithm is the most widely used, as presented in Figure 6. With a small number of selected genes, the genetic algorithm obtains the best accuracy.
- 5) The most-used classifier in the literature review is the SVM, followed by KNN and NB, as shown in Figure 7. The classifier that gave the highest accuracy while selecting a low number of genes was the SVM. Moreover, only one study [82] uses CNN as the classifier, but the study did not specify the number of selected genes. However, the SVM gives the worst accuracy in MOBBA-LS [83].
- 6) Hybrid SVM-RFE and BDF give great accuracy, but the number of selected genes is higher than 1000, which is the largest number of genes of any hybrid method.
- 7) Figure 8 presents the most common wrapper methods' meta-heuristic categories. The methods used are swarm based and evolution based so they can produce high accuracy and select fewer genes.
- 8) By looking at [71], we can see that even though the authors used BD, which is considered a good bio-inspired method that has shown promising results in other papers, it selected a huge number of genes. We will try to improve the dragonfly method to select fewer genes.

## VII. CONCLUSION

Hybrid methods have grown in popularity in recent years, this paper present different proposed and described the hybrid feature selection methods (Filter/Wrapper and

Wrapper/Wrapper) methods and compared their performance between each other regarding the number of selected genes and their accuracy. While various models have been proposed to solve the dimensionality issue of microarray gene expression profiles, specifically their accuracy and number of selected genes. we only looked at papers that were published between 2017-2021 that proposed a hybrid feature selection method in order to maintain an appropriate number of papers and focus on only the recent years.

We Found out that GA is the most commonly used wrapper method. In addition, we look at which dataset is commonly used to evaluate the developed methods which is in n order, Colon [90], SRBC [88], Colon [88]. Most commonly used in classifying cancer is the Support Vector Machine (SVM).

## REFERENCES

- [1] *Cancer*. Accessed: Aug. 30, 2021. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/cancer>
- [2] *Global-Cancer-Rates-Could-Increase-by-to-Million-Global Cancer Rates Could Increase by 50% to 15 Million*. Accessed: May 20, 2022. [Online]. Available: <https://www.who.int/news>
- [3] C. Hadley King and A. A. Sinha, "Gene expression profile analysis by DNA microarrays: Promise and pitfalls," *J. Amer. Med. Assoc.*, vol. 286, no. 18, pp. 2280–2288, 2001.
- [4] S. Alagukumar and R. Lawrance, "Classification of microarray gene expression data using associative classification," in *Proc. Int. Conf. Comput. Technol. Intell. Data Eng. (ICCTIDE)*, Jan. 2016, pp. 1–8.
- [5] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, *Feature Extraction: Foundations and Applications* (Studies in Fuzziness and Soft Computing). Berlin, Germany: Springer-Verlag, 2008.
- [6] Y. Wang, F. S. Makedon, J. C. Ford, and J. Pearlman, "HykGene: A hybrid approach for selecting marker genes for phenotype classification using microarray gene expression data," *Bioinformatics*, vol. 21, no. 8, pp. 1530–1537, Apr. 2005.
- [7] V. Bolón-Canedo, N. Sánchez-Marroño, 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.
- [8] Y. Saeyns, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.
- [9] D. Roobaert, G. Karakoulas, and V. Nitesh Chawla, "Information gain, correlation and support vector machines," in *Feature Extraction: Foundations and Applications* (Studies in Fuzziness and Soft Computing), I. Guyon, M. Nikravesh, S. Gunn, and L. A. Zadeh, Eds. Berlin, Germany: Springer, 2006, pp. 463–470.
- [10] S. R. Shinde and P. S. Bhadikar, "A genetic algorithm, information gain and artificial neural network based approach for hypertension diagnosis," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, Aug. 2016, pp. 1–7.
- [11] *What is Mutual Information? Quantdare*. Accessed: Mar. 31, 2021. [Online]. Available: <https://quantdare.com/what-is-mutual-information/>
- [12] G. Bontempi and E. Patrick Meyer, "Causal filter selection in microarray data," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 95–102.
- [13] 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.
- [14] Z. Zhao, R. Anand, and M. Wang, "Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform," in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics (DSAA)*, Oct. 2019, pp. 442–452.
- [15] L. Jiang, "Learning random forests for ranking," *Frontiers Comput. Sci. China*, vol. 5, no. 1, pp. 79–86, Mar. 2011.
- [16] K. S. Mohsen and A. T. Sadiq, "Random forest algorithm using accuracy-based ranking," *J. Comput. Theor. Nanoscience*, vol. 16, no. 3, pp. 1039–1045, Mar. 2019.
- [17] L. Yu and H. Liu, "Feature selection for high-dimensional data," in *Proc. 20th Int. Conf. Mach. Learn. (ICML)*, Washington, DC, USA, 2003, pp. 856–863.
- [18] 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. Electr. Eng. Boumerdes (ICEE-B)*, Oct. 2017, pp. 1–6.
- [19] C. Li and J. Xu, "Feature selection with the Fisher score followed by the maximal clique centrality algorithm can accurately identify the hub genes of hepatocellular carcinoma," *Sci. Rep.*, vol. 9, no. 1, p. 17283, Dec. 2019.
- [20] Y.-W. Chen and C.-J. Lin, "Combining SVMs with various feature selection strategies," in *Feature Extraction: Foundations and Applications* (Studies in Fuzziness and Soft Computing), I. Guyon, M. Nikravesh, S. Gunn, L. A. Zadeh, Eds. Berlin, Germany: Springer, 2006, pp. 315–324.
- [21] J. Zheng, H. Zhu, F. Chang, and Y. Liu, "An improved relief feature selection algorithm based on monte-carlo tree search," *Syst. Sci. Control Eng.*, vol. 7, no. 1, pp. 304–310, Jan. 2019, doi: [10.1080/21642583.2019.1661312](https://doi.org/10.1080/21642583.2019.1661312).
- [22] K. Kira and A. Larry Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Proc. 10th Nat. Conf. Artif. Intell.*, 1992, pp. 129–134.
- [23] P. Ngatchou, A. Zarei, and A. El-Sharkawi, "Pareto multi objective optimization," in *Proc. 13th Int. Conf. Intell. Syst. Appl. Power Syst.*, Nov. 2005, pp. 84–91.
- [24] J. Pirgazi, M. Alimoradi, T. E. Abharian, and M. H. Olyae, "An efficient hybrid filter-wrapper Metaheuristic-based gene selection method for high dimensional datasets," *Sci. Rep.*, vol. 9, no. 1, p. 18580, Dec. 2019.
- [25] J. Suto, S. Oniga, and P. P. Sitar, "Comparison of wrapper and filter feature selection algorithms on human activity recognition," in *Proc. 6th Int. Conf. Comput. Commun. Control (ICCCC)*, May 2016, pp. 124–129.
- [26] T. Nguyen, G. Nguyen, and B. M. Nguyen, "EO-CNN: An enhanced CNN model trained by equilibrium optimization for traffic transportation prediction," *Proc. Comput. Sci.*, vol. 176, pp. 800–809, Jan. 2020.
- [27] S. Katoch, S. S. Chauhan, and V. Kumar, "A review on genetic algorithm: Past, present, and future," *Multimedia Tools Appl.*, vol. 80, pp. 8091–8126, Oct. 2020.
- [28] J. McCall, "Genetic algorithms for modelling and optimisation," *J. Comput. Appl. Math.*, vol. 184, no. 1, pp. 205–222, Dec. 2005.
- [29] Z. A. A. Alyasseri, A. T. Khader, M. A. Al-Betar, A. M. Awadallah, and X.-S. Yang, "Variants of the flower pollination algorithm: A review," in *Nature-Inspired Algorithms and Applied Optimization* (Studies in Computational Intelligence), X.-S. Yang, Ed. Berlin, Germany: Springer, 2018, pp. 91–118.
- [30] W. Cui and Y. He, "Biological flower pollination algorithm with orthogonal learning strategy and catfish effect mechanism for global optimization problems," *Math. Problems Eng.*, vol. 2018, pp. 1–16, Jan. 2018.
- [31] D. Karaboga *et al.*, "An idea based on honey bee swarm for numerical optimization," Dept. Comput. Eng., Eng. Fac., Erciyes Univ., Tech. Rep. tr06, 2005.
- [32] X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," in *Proc. World Congr. Nature Biologically Inspired Comput. (NaBIC)*, Dec. 2009, pp. 210–214.
- [33] M. M. Mafarja, D. Eleyan, I. Jaber, A. Hammouri, and S. Mirjalili, "Binary dragonfly algorithm for feature selection," in *Proc. Int. Conf. New Trends Comput. Sci. (ICTCS)*, Oct. 2017, pp. 12–17.
- [34] C. M. Rahman and T. A. Rashid, "Dragonfly algorithm and its applications in applied science survey," *Comput. Intell. Neurosci.*, vol. 2019, pp. 1–21, Dec. 2019.
- [35] S. Mirjalili, "Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm," *Knowl.-Based Syst.*, vol. 89, pp. 228–249, Nov. 2015.
- [36] I. Ahmad, "Feature selection using particle swarm optimization in intrusion detection," *Int. J. Distrib. Sensor Netw.*, vol. 11, no. 10, Oct. 2015, Art. no. 806954.
- [37] F. Ardjani, K. Sadouni, and M. Benyettou, "Optimization of SVM MultiClass by particle swarm (PSO-SVM)," in *Proc. 2nd Int. Workshop Database Technol. Appl.*, Nov. 2010, pp. 1–4.
- [38] H. Xu, S. Yu, J. Chen, and X. Zuo, "An improved firefly algorithm for feature selection in classification," *Wireless Pers. Commun.*, vol. 102, no. 4, pp. 2823–2834, Oct. 2018.
- [39] E. M. Mashhour, E. M. F. El Houbay, K. T. Wassif, and A. I. Salah, "Feature selection approach based on firefly algorithm and chi-square," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 4, p. 2338, Aug. 2018.

- [40] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," in *Nature Inspired Cooperative Strategies for Optimization* (Studies in Computational Intelligence), J. R. González, D. A. Pelta, C. Cruz, G. Terrazas, and N. Krasnogor, Eds. Berlin, Germany: Springer, 2010, pp. 65–74.
- [41] J. Huang and Y. Ma, "Bat algorithm based on an integration strategy and Gaussian distribution," *Math. Problems Eng.*, vol. 2020, pp. 1–22, Oct. 2020.
- [42] Y. Tamura, T. Sakiyama, and I. Arizono, "Ant colony optimization using common social information and self-memory," *Complexity*, vol. 2021, pp. 1–7, Jan. 2021.
- [43] M. Dorigo, M. Birattari, and T. Stuetzle, "Ant colony optimization: Artificial ants as a computational intelligence technique," *IEEE Comput. Intell. Mag.*, vol. 1, no. 4, pp. 28–39, Jan. 2006.
- [44] R. V. Rao and V. Patel, "An improved teaching-learning-based optimization algorithm for solving unconstrained optimization problems," *Scientia Iranica*, vol. 20, no. 3, pp. 710–720, 2013.
- [45] Y. Su, K. Qi, C. Di, Y. Ma, and S. Li, "Learning automata based feature selection for network traffic intrusion detection," in *Proc. IEEE 3rd Int. Conf. Data Sci. Cyberspace (DSC)*, Jun. 2018, pp. 622–627.
- [46] R. Soto, B. Crawford, R. Olivares, C. Taramasco, I. Figueroa, Á. Gómez, C. Castro, and F. Paredes, "Adaptive black hole algorithm for solving the set covering problem," *Math. Problems Eng.*, vol. 2018, pp. 1–23, Oct. 2018.
- [47] S. Nagpal, S. Arora, S. Dey, and Shreya, "Feature selection using gravitational search algorithm for biomedical data," *Proc. Comput. Sci.*, vol. 115, pp. 258–265, Jan. 2017.
- [48] J. P. Papa, A. Pagnin, S. A. Schellini, A. Spadotto, R. C. Guido, M. Ponti, G. Chiachia, and A. X. Falcao, "Feature selection through gravitational search algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 2052–2055.
- [49] 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.
- [50] A. Sampathkumar, R. Rastogi, S. Arukonda, A. Shankar, S. Kautish, and M. Sivaram, "An efficient hybrid methodology for detection of cancer-causing gene using CSC for micro array data," *J. Ambient Intell. Hum. Comput.*, vol. 11, no. 11, pp. 4743–4751, Nov. 2020.
- [51] O. Aouedi, K. Piamrat, and D. Bagadthey, "A semi-supervised stacked autoencoder approach for network traffic classification," in *Proc. IEEE 28th Int. Conf. Neww. Protocols (ICNP)*, Oct. 2020, pp. 1–6.
- [52] G. Liu, H. Bao, and B. Han, "A stacked autoencoder-based deep neural network for achieving gearbox fault diagnosis," *Math. Problems Eng.*, vol. 2018, pp. 1–10, Jul. 2018.
- [53] *Binary and Multiclass Classification in Machine Learning | Analytics Steps*. Accessed: Oct. 9, 2021. [Online]. Available: <https://www.analyticssteps.com/blogs/binary-and-multiclass-classification-machine-learning>
- [54] *A Complete Guide to Understand Classification in Machine Learning*. Accessed: Sep. 9, 2021. [Online]. Available: <https://www.analyticsvidhya.com/blog/2021/09/a-complete-guide-to-understand-classification-in-machine-learning/>
- [55] P. A. Telnoni, R. Budiawan, and M. Qana'a, "Comparison of machine learning classification method on text-based case in Twitter," in *Proc. Int. Conf. ICT Smart Soc. (ICISS)*, Nov. 2019, pp. 1–5.
- [56] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods: Support Vector Learning*, 1999.
- [57] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, "Support vector machine (SVM)," in *Introduction to Data Mining*, 2nd ed. London, U.K.: Pearson, 2006, pp. 256–276.
- [58] K.-H. Liu, M. Tong, S.-T. Xie, and V. T. Y. Ng, "Genetic programming based ensemble system for microarray data classification," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–11, 2015.
- [59] L. Guo, D. Rivero, J. Dorado, C. R. Munteanu, and A. Pazos, "Automatic feature extraction using genetic programming: An application to epileptic EEG classification," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 10425–10436, Aug. 2011.
- [60] A. Wang, N. An, G. Chen, L. Li, and G. Alterovitz, "Accelerating incremental wrapper based gene selection with K-nearest-neighbor," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2014, pp. 21–23.
- [61] M. Parsian, *Data Algorithms: Recipes for Scaling Up With Hadoop Spark*. Newton, MA, USA: O'Reilly Media, 2015.
- [62] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull. Math. Biophys.*, vol. 5, no. 4, pp. 115–133, 1943.
- [63] M. Bardamova, A. Konev, I. Hodashinsky, and A. Shelupanov, "A fuzzy classifier with feature selection based on the gravitational search algorithm," *Symmetry*, vol. 10, no. 11, p. 609, Nov. 2018.
- [64] M. A. Hambali, T. O. Oladele, and K. S. Adewole, "Microarray cancer feature selection: Review, challenges and research directions," *Int. J. Cognit. Comput. Eng.*, vol. 1, pp. 78–97, Jun. 2020.
- [65] 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.
- [66] E. Pashaei and E. Pashaei, "Gene selection using intelligent dynamic genetic algorithm and random forest," in *Proc. 11th Int. Conf. Electr. Electron. Eng. (ELECO)*, Nov. 2019, pp. 470–474.
- [67] P. Tumuluru, "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.
- [68] H. M. Alshamlan, G. H. Badr, and Y. A. Alohal, "Genetic bee colony (GBC) algorithm: A new gene selection method for microarray cancer classification," *Comput. Biol. Chem.*, vol. 56, pp. 49–60, Jun. 2015.
- [69] E. Pashaei and E. Pashaei, "Gene selection for cancer classification using a new hybrid of binary black hole algorithm," in *Proc. 28th Signal Process. Commun. Appl. Conf. (SIU)*, Oct. 2020, pp. 1–4.
- [70] 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.
- [71] S. A. Medjahed, T. A. Saadi, A. Benyettou, and M. Ouali, "Kernel-based learning and feature selection analysis for cancer diagnosis," *Appl. Soft Comput.*, vol. 51, pp. 39–48, Feb. 2017.
- [72] M. J. Rani and D. Devaraj, "Two-stage hybrid gene selection using mutual information and genetic algorithm for cancer data classification," *J. Med. Syst.*, vol. 43, no. 8, p. 235, Aug. 2019.
- [73] H. Motieghader, A. Najafi, B. Sadeghi, and A. Masoudi-Nejad, "A hybrid gene selection algorithm for microarray cancer classification using genetic algorithm and learning automata," *Informat. Med. Unlocked*, vol. 9, pp. 246–254, 2017.
- [74] A. K. Shukla, P. Singh, and M. Vardhan, "A hybrid gene selection method for microarray recognition," *Biocybernetics Biomed. Eng.*, vol. 38, no. 4, pp. 975–991, 2018.
- [75] M. A. Al-Betar, O. A. Alomari, and S. M. Abu-Romman, "A TRIZ-inspired bat algorithm for gene selection in cancer classification," *Genomics*, vol. 112, no. 1, pp. 114–126, Jan. 2020.
- [76] J. Ge, X. Zhang, G. Liu, and Y. Sun, "A novel feature selection algorithm based on artificial bee colony algorithm and genetic algorithm," in *Proc. IEEE Int. Conf. Power, Intell. Comput. Syst. (ICPICS)*, Jul. 2019, pp. 131–135.
- [77] N. Almgren and H. M. Alshamlan, "New bio-marker gene discovery algorithms for cancer gene expression profile," *IEEE Access*, vol. 7, pp. 136907–136913, 2019.
- [78] A. K. Shukla, P. Singh, and M. Vardhan, "A new hybrid feature subset selection framework based on binary genetic algorithm and information theory," *Int. J. Comput. Intell. Appl.*, vol. 18, no. 3, Sep. 2019, Art. no. 1950020.
- [79] A. K. Shukla, P. Singh, and M. Vardhan, "Gene selection for cancer types classification using novel hybrid metaheuristics approach," *Swarm Evol. Comput.*, vol. 54, May 2020, Art. no. 100661.
- [80] H. Djellali, A. Djebbar, N. G. Zine, and N. Azizi, "Hybrid artificial bees colony and particle swarm on feature selection," in *Computational Intelligence and Its Applications* (Advances in Information and Communication Technology), A. Amine, M. Mouhoub, O. A. Mohamed, and B. Djebbar, Eds. Cham, Switzerland: Springer, 2018, pp. 93–105.
- [81] O. A. Alomari, A. T. Khader, M. A. Al-Betar, and Z. A. A. Alyasseri, "A hybrid filter-wrapper gene selection method for cancer classification," in *Proc. 2nd Int. Conf. BioSignal Anal., Process. Syst. (ICBAPS)*, Jul. 2018, pp. 113–118.
- [82] S. Kilicarslan, K. Adem, and M. Celik, "Diagnosis and classification of cancer using hybrid model based on ReliefF and convolutional neural network," *Med. Hypotheses*, vol. 137, Apr. 2020, Art. no. 109577.
- [83] 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.
- [84] S. A. A. Vijay and P. GaneshKumar, "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, pp. 1–12, Apr. 2018.

- [85] A. Dabba, A. Tari, S. Meftali, and R. Mokhtari, "Gene selection and classification of microarray data method based on mutual information and moth flame algorithm," *Expert Syst. Appl.*, vol. 166, Mar. 2021, Art. no. 114012.
- [86] C. Yan, X. Kang, M. Li, and J. Wang, "A novel feature selection method on mutual information and improved gravitational search algorithm for high dimensional biomedical data," in *Proc. 13th Int. Conf. Comput. Autom. Eng. (ICCAE)*, Mar. 2021, pp. 24–30.
- [87] R. A. Musheer, C. K. Verma, and N. Srivastava, "Novel machine learning approach for classification of high-dimensional microarray data," *Soft Comput.*, vol. 23, no. 24, pp. 13409–13421, Dec. 2019.
- [88] O. A. Alomari, S. N. Makhadmeh, M. A. Al-Betar, Z. A. A. Alyasseri, I. A. Doush, A. K. Abasi, M. A. Awadallah, and R. A. Zitar, "Gene selection for microarray data classification based on gray wolf optimizer enhanced with TRIZ-inspired operators," *Knowl.-Based Syst.*, vol. 223, Jul. 2021, Art. no. 107034.
- [89] R. Dash, "An adaptive harmony search approach for gene selection and classification of high dimensional medical data," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 33, no. 2, pp. 195–207, Feb. 2021.
- [90] R. Aziz, C. K. Verma, and N. Srivastava, "Artificial neural network classification of high dimensional data with novel optimization approach of dimension reduction," *Ann. Data Sci.*, vol. 5, no. 4, pp. 615–635, Dec. 2018.
- [91] S. K. Baliarsingh, K. Muhammad, and S. Bakshi, "SARA: A memetic algorithm for high-dimensional biomedical data," *Appl. Soft Comput.*, vol. 101, Mar. 2021, Art. no. 107009.
- [92] 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, 1999.
- [93] *UCI Machine Learning Repository: Data Sets*. Accessed: Nov. 28, 2021. [Online]. Available: <https://archive.ics.uci.edu/ml/datasets.php>
- [94] A. Statnikov, I. Tsamardinos, Y. Dosbayev, and C. F. Aliferis, "GEMS: A system for automated cancer diagnosis and biomarker discovery from microarray gene expression data," *Int. J. Med. Informat.*, vol. 74, nos. 7–8, pp. 491–503, Aug. 2005.
- [95] 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, 1999.
- [96] D. Singh, G. P. Febbo, K. Ross, G. D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, V. Anthony D'Amico, P. J. Richie, S. E. Lander, M. Loda, W. P. Kantoff, R. T. Golub, and R. W. Sellers, "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol. 1, no. 2, pp. 203–209, 2002.
- [97] G. J. Gordon, R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer Res.*, vol. 62, pp. 4963–4967, Sep. 2002.
- [98] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, no. 15, pp. 2429–2437, 2004.
- [99] "Dragonfly algorithm: A new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems," *Neural Comput. Appl.*, vol. 27, no. 4, pp. 1053–1073, 2016, doi: [10.1007/s00521-015-1920-1](https://doi.org/10.1007/s00521-015-1920-1).
- [100] 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.
- [101] F. E. Petricoin, M. A. Ardekani, A. B. Hitt, J. P. Levine, A. V. Fusaro, M. S. Steinberg, B. G. Mills, C. Simone, A. D. Fishman, C. E. Kohn, and A. L. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet*, vol. 359, no. 9306, pp. 572–577, Feb. 2002.
- [102] S. Kar, K. D. Sharma, and M. Maitra, "Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 612–627, Jan. 2015.
- [103] K.-H. Chen, K.-J. Wang, M.-L. Tsai, K.-M. Wang, A. M. Adrian, W.-C. Cheng, T.-S. Yang, N.-C. Teng, K.-P. Tan, and K.-S. Chang, "Gene selection for cancer identification: A decision tree model empowered by particle swarm optimization algorithm," *BMC Bioinf.*, vol. 15, no. 1, p. 49, Dec. 2014.
- [104] J. L. van 't Veer, H. Dai, J. M. van de Vijver, D. Y. He, A. M. A. Hart, M. Mao, L. H. Peterse, K. V. D. Kooy, J. M. Marton, T. A. Witteveen, J. G. Schreiber, M. R. Kerkhoven, C. Roberts, S. P. Linsley, R. Bernards, and H. S. Friend, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.
- [105] *UCI Machine Learning Repository: Data Sets*. Accessed: Nov. 29, 2021. [Online]. Available: <http://csse.szu.edu.cn/staff/zhuzx/Datasets.html>
- [106] *Kent Ridge Biomedical Data Set Repository*. Accessed: Nov. 27, 2021. [Online]. Available: <https://leo.ugr.es/elvira/DBCRepository/>
- [107] G. D. Beer, L. R. S. Kardia, C.-C. Huang, J. T. Giordano, M. A. Levin, E. D. Misek, L. Lin, G. Chen, G. T. Gharib, G. D. Thomas, L. M. Lizyness, R. Kuick, S. Hayasaka, M. G. J. Taylor, D. M. Iannettoni, B. M. 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.
- [108] J. Khan, S. J. Wei, M. Ringnér, H. L. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, R. C. Antonescu, C. Peterson, and S. P. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Med.*, vol. 7, no. 6, pp. 673–679, 2001.
- [109] 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.
- [110] S. L. Pomeroy *et al.*, "Prediction of central nervous system embryonal tumour outcome based on gene expression," *Nature*, vol. 415, no. 6870, pp. 436–442, Jan. 2002.
- [111] P. Pulkkinen and H. Koivisto, "Identification of interpretable and accurate fuzzy classifiers and function estimators with hybrid methods," *Appl. Soft Comput.*, vol. 7, no. 2, pp. 520–533, 2007.
- [112] Z. Zhu, Y.-S. Ong, and M. Dash, "Markov blanket-embedded genetic algorithm for gene selection," *Pattern Recognit.*, vol. 40, no. 11, pp. 3236–3248, Nov. 2007.
- [113] *Home—Geo—NCBI*. Accessed: Jan. 24, 2021. [Online]. Available: <https://www.ncbi.nlm.nih.gov/geo/>
- [114] C. L. Nutt, D. R. Mani, R. A. Betensky, P. Tamayo, J. G. Cairncross, C. Ladd, U. Pohl, C. Hartmann, M. E. McLaughlin, T. T. Batchelor, and P. M. Black, "Gene expression-based classification of malignant gliomas correlates better with survival than histological classification," *Cancer Res.*, vol. 63, no. 7, pp. 1602–1607, 2003.
- [115] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, and M. Loda, "Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses," in *Proc. Nat. Acad. Sci. USA*, vol. 98, pp. 13790–13795, Nov. 2001.

**HALAH ALMAZRUA** received the bachelor's and master's degrees in information technology from King Saud University, in 2018 and 2022, respectively.

She is currently working as a DevOps Engineer and a Researcher. Her research interests include computer science in the areas of bioinformatics, microarray data analysis, cloud robotics, and artificial intelligence.

**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 Laboratory, Department of Mechanical Engineering, Massachusetts Institute of Technology (MIT), Cambridge, USA. In 2018, she was the Leader of the Data Science for Global Health Track at MIT Hacking Medicine, Riyadh, Saudi Arabia. She is currently an Assistant Professor with the Department of Information Technology, 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 cancer biomarkers from genomic data. All these algorithms have been published in high-impact journals. Her research interests include bioinformatics, especially how artificial intelligence techniques and machine learning approaches can be applied to the analysis of biological data.

...