Available online freely at www.isisn.org

Bioscience Research

Print ISSN: 1811-9506 Online ISSN: 2218-3973

Journal by Innovative Scientific Information & Services Network

RESEARCH ARTICLE BIOSCIENCE RESEARCH, 2019 16(3): 3139-3154. **OPEN ACCESS**

A new filter-based Gene selection method based on dragonfly optimization and correlation-based feature selection

Mohamed Ghoneimy[1], Emad Nabil[2,3], Amr Badr[2], Sherif F. El-Khamisy[4]

[1]Faculty of Information Technology, MUST University, 6th of October City, Giza, **Egypt.**
[2]Faculty of Computers and Artificial Intelligence, Cairo University, Giza, **Egypt**.
[3]Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah, **Saudi Arabia**.
[4]Center for Genomics, Helmy Institute, Zewail City of Science and Technology, Giza 12588, **Egypt.**

***Correspondence:** Mohamed.Ghoneimy@must.edu.eg **Accepted: 0**0 July 2019 **Published online:** 12Sep 2019

Cancer Diagnosis is considered one of microarray data's most developing applications. But the classification of cancer using microarray data stills a hard problem, this is because of the microarray data consists of a massive number of genes and a small number of cases. In order to tackle this problem a gene selection method must be used which improves the accuracy of classification. A new filter-based gene selection method is proposed in this paper. The proposed method merges the Dragonfly algorithm and the correlation-based feature selection, this is to reduce the redundancy between the genes selected and increase the relevance between the selected genes and the decision. Our proposed method is compared with nine famous feature selection methods. The experiments in this paper are applied to five widely used public microarray datasets. The used evaluation criterion of the selected features is the average accuracy of classification using three different classifiers, which are support vector machine, naïve Bayes, and decision tree. Experimental results demonstrate that our proposed method is efficient and performs better than the other nine methods used in the experiment. It also shows that the proposed method can be used with anyone of the three classifiers included in our study to obtain an efficient automatic cancer diagnostic system.

Keywords: Feature Selection; Filter Approach; Gene Selection; Microarray Data; DragonFly Optimization; Correlation Coefficient

INTRODUCTION

Cancer is a very dangerous disease which occurs from an uncontrolled division of cells. Its diagnosis is a very complicated task. It is now considered the world's deadliest disease. One of the most significant factors in increasing survival rates is an early diagnosis. Using classifier systems for cancer diagnosis can aid specialists in making an insightful and more confident diagnosis.

In the medical domain, the microarray is used to generate molecular profiles of normal and diseased tissues of patients. Using these profiles can help experts in understanding several diseases. These profiles also are very helpful in both diagnosis and prognosis process. The original microarray data is images which converted into matrices. In the converted matrices, the rows represent the genes while the columns represent the samples. The value in each field represents the expression level of a certain gene in a certain case (Golub et al., 1999). In the course of the most recent two decades, many machine learning methods applied on the

microarray datasets (Shyamsundar et al., 2005; Rubio-Escudero et al. 2008; Rapaport et al. 2007; Gutiérrez-Avilés et al., 2014). Many recent publications proposed gene selection or classification methods that work on the microarray datasets (Salem et al., 2017; Dashtban & Balafar 2017; Nguyen et al., 2015; Elyasigomari et al. 2017).

The curse of dimensionality is the main disadvantage of gene expression data. The number of genes is somewhere in the range of 20,000 and 30,000 while the number of samples is under 150. Repeated and unrelated features are the biggest issue of dimensionality problem. Feature/gene selection is a popular preprocessing method in microarray domain that selects a subset of genes which are rich with information from the initial gene set. Using this technique increases the classification performance as long as reduces the computational costs for any diagnostic system (Tabakhi et al., 2014; Li et al., 2013; Niijima and Okuno 2009; Cai et al., 2009; Liu and Yu 2005).

In gene selection, nature-inspired and evolutionary optimization algorithms are very useful. In this work, dragonfly optimization (DF) is used for gene selection. DF is an optimization technique based on swarm intelligence. The DF algorithm's main inspiration is the static and dynamic swarming behaviors of dragonflies in the natural world. The algorithm solves optimization problems by modeling dragonfly's social interaction in the search for food , navigation and enemy avoidance(Mirjalili 2016).

Therefore, the goal of this work is to design a new gene selection method that merges the DF algorithm's good performance with the filter approach's computational efficiency.

In this paper, A new filter-based gene selection method for microarray data is proposed which name is DOC-FS (Dragonfly Optimization and Correlation-based Features Selection). DOC-FS is a repeated improvement process while at each repetition the group of dragonflies selects subsets of features/genes. Thereafter, DOC-FS measures the fitness of the founded subsets using CFS (Correlation-based Features Selection) without using any classifier. Eventually, DOC-FS chooses the best subset of features/genes as the selected gene set. The selected genes can be used with a classifier to build a diagnostic system. They also can be used in further biological research to determine the biological relevance between cancer and the selected genes.

The remainder of this manuscript is organized as follows. In the second section, the related and recent work on the use of feature selection methods in cancer prediction using gene expression data. The third section presents DOC-FS method while the fourth section provides the experimental results, statistical analysis. The fifth section discusses the obtained results and Finally, the sixth section presents the conclusion of this study and offers some overall perspective.

Related Works

This section presents and discusses the related works to gene/feature selection microarray data. There are generally four categories of methods of gene selection which are wrapper approach, filter approach, hybrid approach, and embedded approach (Leung & Hung 2010; Saeys et al. 2007; Bolón-Canedo et al., 2014; Li et al., 2013; Inza et al., 2004) as shown in Figure 1. In the case of the filter approach, the relevance of genes is measured by using the statistical properties of the data which doesn't need to use any classifier.

Many strategies for measuring the relation of genes exist, such as univariate and multivariate (Lazar et al. 2012; Tabakhi et al., 2014; Peng et al., 2005; Golub et al., 1999). In univariate strategies, in the beginning, the feature selection method measures and sorts the genes according to a certain criterion, then the top genes according to the fitness are chosen as the best subset of features/genes. Different criteria are used in univariate strategies including information gain (Raileanu and Stoffel 2004), mutual information (Cai et al., 2009), Laplacian score (LS) (Liao et al. 2014; He et al., 2006), term variance (TV) (sergoios Theodoridis 2008), Signal-to-noise ratio (Golub et al., 1999). The advantages of univariate-based methods are their high speed and efficiency. On the other hand, they ignore the dependencies between selected genes which lead to lower classification accuracy. The multivariate strategy considers the correlation between selected genes which allow the multivariate methods getting higher classification accuracy than univariate methods. Different multivariate methods are exists such as mutual correlation (MC) (Haindl et al., 2006; Ghazavi & Liao 2008), unsupervised feature selection based on the ant colony optimization method (UFSACO) (Tabakhi et al., 2014), random subspace method (RSM) (Bertoni et al., 2005; Lai, Reinders & Wessels 2006; Li & Zhao 2009), minimal-redundancy-

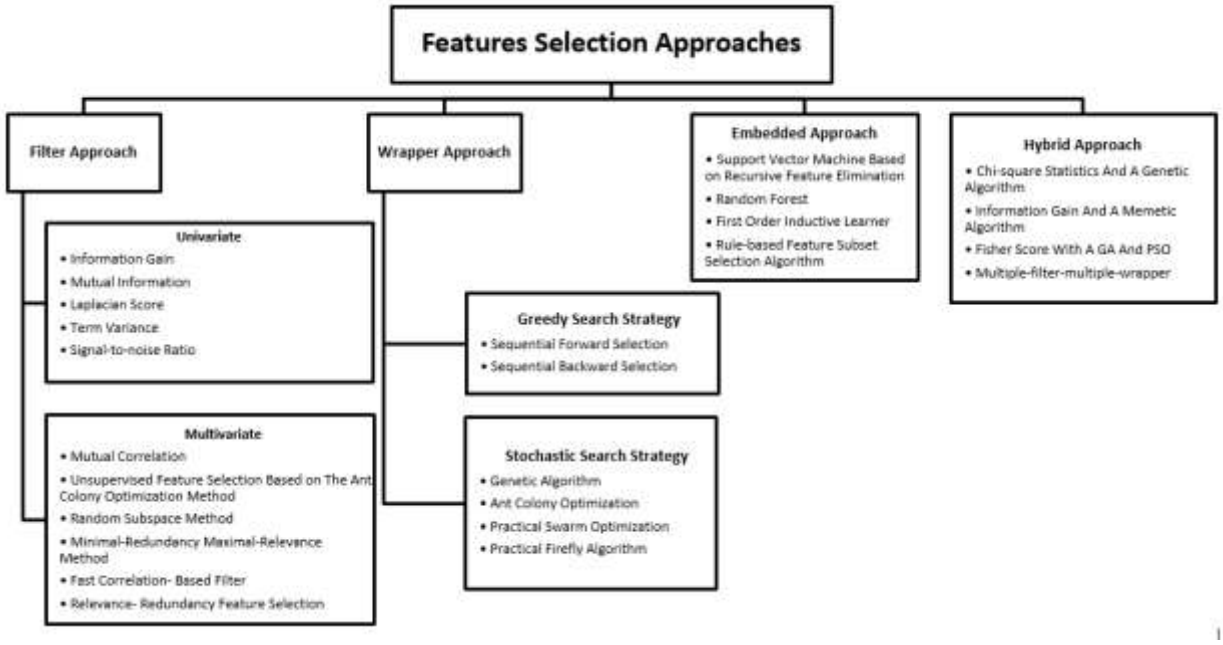

Figure 1; Approaches of Feature selection

maximal-relevance method (MRMR) (Peng et al., 2005; Ding & Peng 2005), fast correlation-based filter (FCBF) (Yu and Liu 2004; Yu and Liu 2003), and Relevance- redundancy feature selection (RRFS) (Ferreira and Figueiredo 2012b; Ferreira and Figueiredo 2012a). The search algorithms used in the multivariate method, search for the top subset of features in one iteration, so, this kind of methods may readily be stuck into a local optimum.

A certain classifier is used in the wrapper appr oach to assessing the subset of selected features. Furthermore, the process of search is guided by the chosen classifier's accuracy. Stochastic and greedy search strategy are two basic search strategies for the wrapper approach (Gheyas & Smith 2010; Saeys et al., 2007). The greedy search strategy always uses one of two methods which are sequential forward selection and sequential backward selected (Inza et al., 2004; Inza et al. 2002). Many methods depends on the stochastic search strategy like ant colony optimization (ACO), genetic algorithm (GA) (Li et al., 2013; Kabir et al., 2012; Yu et al., 2009), practical swarm optimization (PSO) (Sahu & Mishra 2012; Martinez et al., 2010), the firefly algorithm (Srivastava et al. 2013), and the dragonfly algorithm (DF) (Mirjalili 2016). Due to the use of a given classifier, the average classification precision of the wrapper approach is considered higher than the filter approach. On the other side, the drawback of wrapper methods is the long computation time, particularly with the microarray datasets. Besides that, the wrapper approach is considered a black box that suffers from the lake of interpretation.

The embedded approach uses a trained learning model with an original feature set to determinethe criterion for the measurement of gen e rank values. Many methods based on embedded approach such as support vector machine based on recursive feature elimination (SVM-RFE) (Guyon et al., 2002), random forest (RF) (Ramón and De Andres 2006), and the first-order inductive learner (FOIL) rule-based feature subset selection algorithm (Wang et al., 2013). The interaction with the learning model is the merit of the embedded approach, however, using all features in the set to train a classifier, consumes a long time particularly with the microarray datasets.

The hybrid methods merge the benefits of wrapper methods and filter methods. The hybrid methods start with selecting features subset using the filter method, after that the wrapper method chooses the final gene set. Since the size of genes is reduced, the computation time needed for the wrapper approach becomes acceptable.

Many methods based on hybrid approach such as chi-square statistics and a GA (Lee and Leu 2011), information gain and a memetic algorithm (Zibakhsh and Abadeh 2013), Fisher score with a GA and PSO (Zhao et al., 2011), and the multiple-filter-multiple-wrapper (MFMW) method (Leung and Hung 2010). The fact that the wrapper and the filter methods are not really merged, and that may cause a worse classification accuracy, is the main weakness of the hybrid approach.

Swarm intelligence-based methods are multi-agent systems in which each artificial agent has a collective attitude. Many swarm intelligence algorithms exist such as ant colony optimization (ACO) (Dorigo & Stützle 2003), PSO (Shi & others 2001), and DF (Mirjalili 2016).Many publications used these methods to tackle the problem of feature selection in many domains like text classification (Aghdam et al., 2009), financial domains (Marinakis et al. 2009), and face recognition (Kanan and Faez 2008). Since most of these methods need a classifier, using them in conjunction with microarray data is not frequent because of the high computational effort. Thus, the filter approach is a preferable approach in the microarray gene selection area.

The DF algorithm is first proposed in (Mirjalili 2016)withquite competitive outcomes compared w ith other recognized literature algorithms such as PSO and gravitational search algorithm (GSA). In (Mafarja et al. 2017) the DF is proposed as a wrapper feature selection guided by KNN (Nearest Neighborhood) classifier as a fitness function. The obtained results demonstrate that the DF algorithm has higher performance than GA and PSO algorithm. To our best knowledge, DF is used in conjunction with microarray datasets only once in (Medjahed et al., 2016). This work proposed a wrapper feature selection method that chooses SVM as a fitness function and DF as a search strategy.

MATERIALS AND METHODS

A new gene selection method based on DF, called DOC-FS, for microarray data classification, is presented in this section as shown in Figure 2. DOC-FS is composed of the DF as a search strategy and CFS as a fitness function. This combination is not proposed before and to our best knowledge, DF is being used for the first time as a search stra tegy for a filter-based feature/gene selection method.

The details of the proposed method's first pha se are outlined in the InfoGain subsection. The Dragonfly Algorithm subsection shows how the DF algorithm is used in conjunction with Correlation-based Feature Selection (CFS) algorithm to form a gene selection method.

InfoGain

The first process in the proposed method is - a univariant strategy - used to select the best genes in term of information gain. The information gain value is calculated by (Equation 1) for each gene. In Equation 1, the entropy of the dataset is computed first then, the dataset is divided into subsets, and the entropy of each subset is calculated. After that, the difference between the former entropy and the weighted sum of the latter ones is returned (Witten & Frank 2005). Finally, only the top genes in terms of InfoGain will be processed in the proposed method's next phase. The selected genes number is chosen by the user of the proposed system.

InfoGain(Class,Attribute) = Entropy(Class) - Entropy(Class | gene) Equation 1

Dragonfly Algorithm

The filter approach chooses features' statistical properties without using any classifier. The newly proposed optimization algorithm DF has been used in the proposed method as a search strategy. DF main inspiration came from the dynamic and static swarming attitude of dragonflies in the natural world as shown in Figure 3. Exploitation and exploration, which are the two main phases of optimization, are planned by modeling the dragonflies' social interaction in navigating, foods searching, and enemies avoidance when swarming statistically or dynamically (Mirjalili 2016). DF has been chosen in the proposed work for three main reasons:

Firstly, DF is a promising algorithm as it's superior to other famous algorithms as mentioned in section 2.

Secondly, to our best knowledge DF algorithm has not been used in filter feature selection approach.

Thirdly, DF is a relatively recent metaheuristic that is more efficient than GA and PSO (Medjahed et al. 2016).

The three above points motivated us to explore the power of DF in tackling the feature selection problem.

Dragonfly is an Odonata family insect. It is one of the smallest predators that hunt most of the small insects. As most of the search strategies, dragonfly algorithm consists of two main phases:

the exploitation and exploration (Mirjalili 2016). The behaviors of swarms are mathematically modeled as listed below:

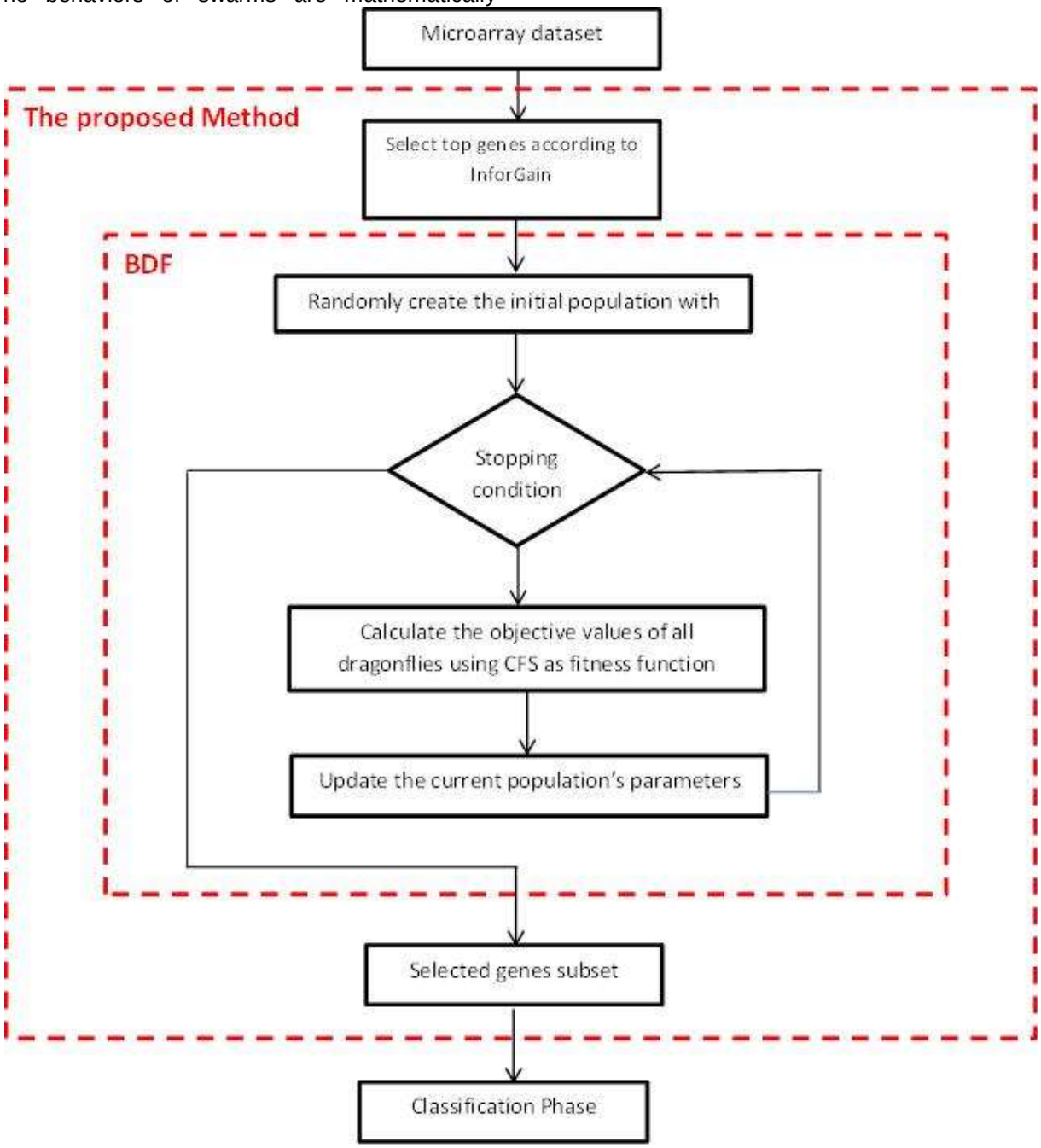

Figure 2; General Schema of DOC-FS method

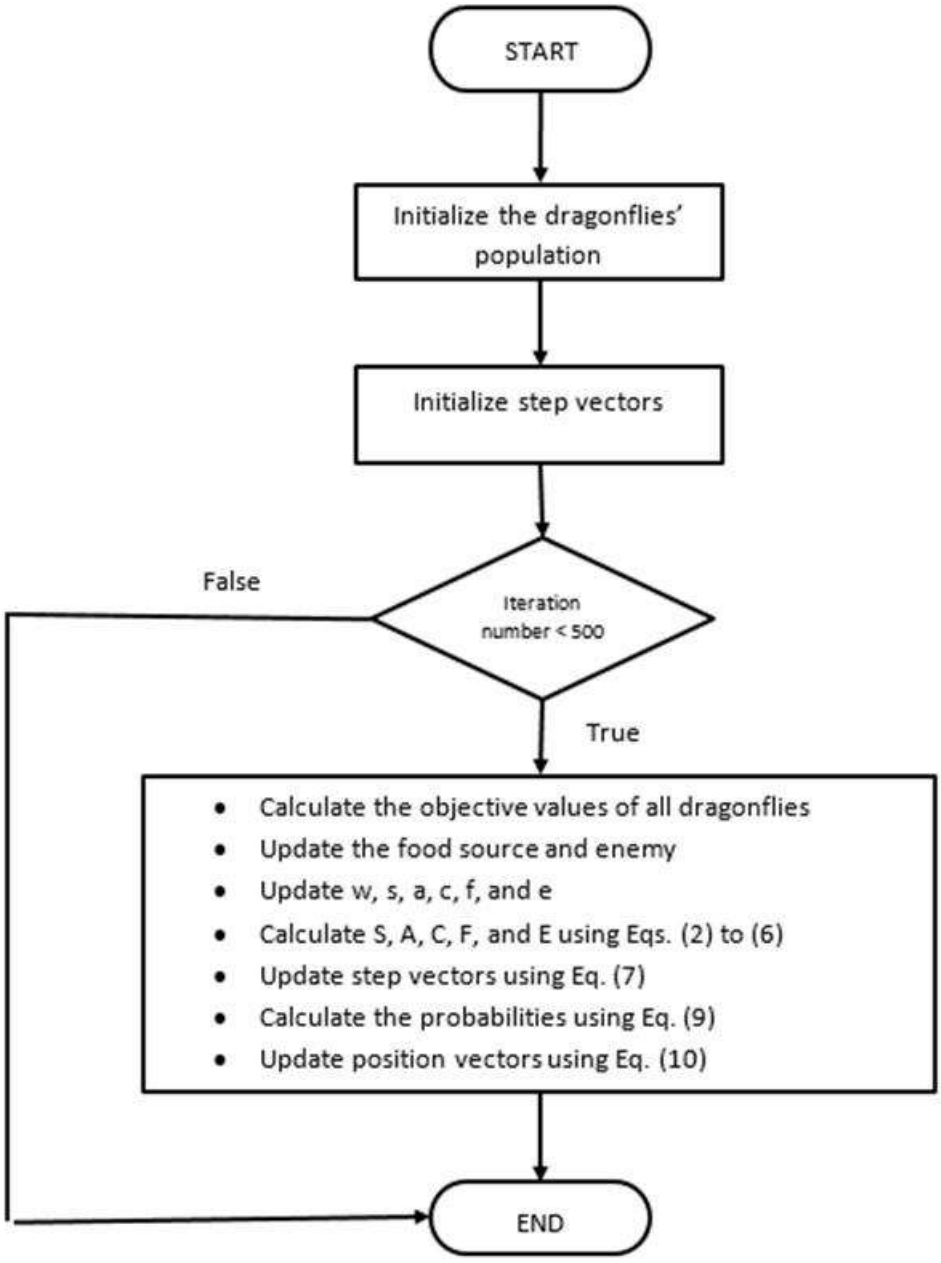

Figure 1 DragonFly Optimization Flowchart

Dragonfly separation is the avoidance of other nei ghborhood dragonflies and is defined as:

$$S_i = -\sum_{j=1}^{n} X_i - X_j \qquad \text{Equation 2}$$

X_i is the current dragonfly's position and X_j is the j^{th} dragonfly's position in the same neighborhood. The number of dragonfly in the neighborhood is n (Mirjalili 2016).

Dragonfly alignment indicates the current dragonfly velocity compared to other dragonflies in the same neighborhood and it can be defined as below:

$$A_i = \frac{\sum_{j=1}^{n} X_j}{n} \qquad \text{Equation 3}$$

X_j is the velocity of the j^{th} dragonfly in the same neighborhood (Mirjalili 2016).
The cohesion *C* of the dragonfly describes the swarm's tendency towards the center of the ex isting neighborhood's mass and can be calculated as below:

$$C_i = \frac{\sum_{j=1}^{n} X_j}{n} - X_i \quad \text{Equation 4}$$

The attraction of the dragonfly *F* simulates the dragonfly's food source attraction and it is defined as:

$$F_i = X^+ - X_i \quad \text{Equation 5}$$

X^+ is the food source's position.
The dragonfly distraction E simulates the outwards of other predators and it's computed as:

$$E_i = X^- + X_i \quad \text{Equation 6}$$

X^- is the enemy's position.
Two vectors are used to represent dragonflies' position in search space and modeling their movements which are position (*X*) and step (Δ*X*). S^i, A^i, C^i, F^i, E^i are used in updating the dragonfly's position as below:

$$\Delta X_{t+1} = (sS_i + aA_i + cC_i + dF_i + eE_i) + w\Delta X_t \quad \text{Equation 7}$$

$$X_{i,t+1} = X_{i,t} + \Delta X_{i,t+1} \quad \text{Equation 8}$$

The parameters *a, s, c, e*, and *d* are the weight of alignment, separation, cohesion, enemy position and food source respectively. t is the iteration number and *w* is inertia weight (Mirjalili 2016).
This algorithm is originally used for solving the continuous optimization problem. The following transfer function has been employed in order to adapt this continuous search algorithm to solve the binary problem. (Mirjalili & Lewis 2013).

$$T(\Delta x) = \left|\frac{\Delta x}{\sqrt{\Delta x^2 + 1}}\right| \quad \text{Equation 9}$$

Calculating the probability of changing position for a dragonfly is achieved by this transfer function. A function to update dragonfly's position in the space of the binary search is presented below:

$$X_{t+1} = \begin{cases} \overline{X_t}, & r < T(\Delta x_{t+1}) \\ X_t, & r \geq T(\Delta x_{t+1}) \end{cases} \quad \text{Equation 10}$$

In the proposed method, all dragonflies are considered in the same swarm and the simulation of exploitation/exploration is achieved by adapting the following factors (*a, s, c, e, f,* and *w*).

Dragonfly Representation

The information about the solution should be stored in the dragonfly, as Figure 4 shows the dragonfly representation. The binary string is the most used encoding format in gene selection problem. In the dragonfly position vector, there are only two values (one/zero) which indicates whether or not a specific gene is selected. While creating the initial population, a random value is generated for the position of each gene. The genes are selected if their positions' values are higher than 0.5, otherwise, they are ignored.

The objective function (CFS)

CFS is a Correlation-based Feature selection algorithm can work with both discrete and continuous problems. It's a heuristic algorithm that calculates the fitness value of a subset of features for DF. Unlike, univariate filter approaches CFS can take into account the interaction between features/genes. CFS takes into account the worth of all features, along with the level of intercorrelation between them, according to the class label's prediction. This algorithm is built on the following assumption:
"Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other."(Mark A Hall 1999)

In the field of test theory, a composite test (individual test average or sum) can be designed by using the same principle to predict an external interest variable. In this case, the features are individual tests that calculate relevance aspects to interest variable (decision variable). The objective function is formalized by (Equation 11):

$$fitness = \frac{kV_{fD}}{\sqrt{k + k(k-1)V_{ff}}} \quad \text{Equation 11}$$

(Equation 11), shows how the fitness value is calculated. V_{fD} is the average feature to decision variable correlation. V_{ff} is the feature to feature intercorrelation average. Actually, (Equation 11) is the Pearson correlation, when all variables were standardized. In (Equation 11), the numerator indicates how predictive a group of features is, while the denominator indicates the level of redundancy between them. The problem of irrelevant features has been tackled as they're going to be bad decision variable predictors. CFS also can handle redundant attributes as they will

be highly correlated with one or more of the other attributes (M. A. Hall 1999).

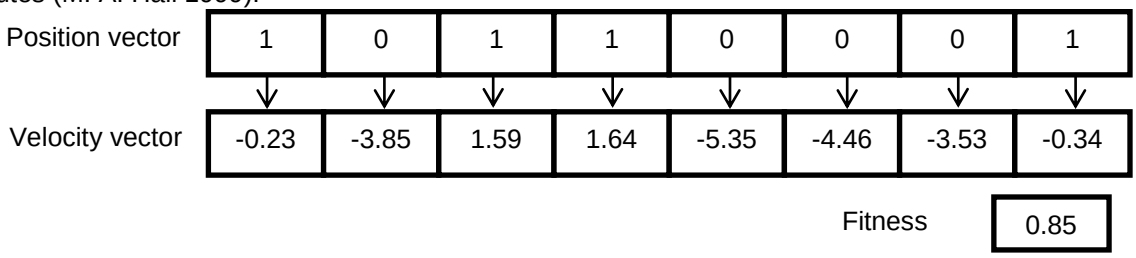

Figure 4 Dragonfly Representation

RESULTS

An empirical assessment of the performance of DOC-FS on five recognized microarray datasets is achieved in this section. DOC-FS was compared with nine extensively used feature selection methods. The proposed method was compared with two univariate filter methods which are term variance (TV), and Laplacian score (LS) (Lai, Reinders, van't Veer, et al., 2006; Liao et al., 2014; sergoios Theodoridis 2008). Both methods can efficiently eliminate genes that are irrelevant.

Four famous multivariate filter methods have been chosen to be compared with the proposed method which are Simplified silhouette filter (SSF) (Covões and Hruschka 2011), Relevance-redundancy feature selection (RRFS) (Ferreira and Figueiredo 2012a; Ferreira and Figueiredo 2012b), random subspace method (RSM) (Ferreira and Figueiredo 2012a; Ferreira and Figueiredo 2012b; Lai, Reinders and Wessels 2006; Li & Zhao 2009), and mutual correlation (MC) (Haindl et al. 2006; Ding and Peng 2005). These methods can detect repeated and unrelated genes.

Furthermore, as DOC-FS is based on swarm intelligence, two swarm-based feature selection methods are selected (UFSACO) (Tabakhi et al. 2014) and microarray gene selection based on ant colony optimization (MGSACO) (Tabakhi et al. 2015).

Eventually, the minimal-redundancy- maximal-relevance method (MRMR) is chosen in the empirical evaluation as a recognized and frequently applied multivariate feature/gene selection method in the literature (Haindl et al., 2006; Ding and Peng 2005).

DOC-FS is supposed to perform well on different classifiers because it's a filter-based feature selection method which doesn't use any classifiers in the gene selection process. Thus, three common and widely used classifiers are selected to assess the efficiency of DOC-FS which is support vector machine (SVM) (Guyon et al., 2002), naïve Bayes (NB) (sergoios Theodoridis 2008), and decision tree (DT) (Quinlan 1986). Many feature/gene selection methods that based on filter approach used the same three classifiers to validate their methods (Lu et al., 2014; Tabakhi et al., 2014; Umamaheswari and Dhivya 2016; Tabakhi et al., 2015).

The WEKA machine learning software library (Hall et al., 2009) has been selected to execute the chosen classifiers. Sequential Minimal Optimization (SMO) is selected for training SVM. In WEKA the choice of SMO was selected. In the SVM classifier, *c* equals 1 - complexity parameter- and the tolerance parameter equals 0.0001. Furthermore, J48 was used as the DT classifier. Regarding the DT classifier, the post-pruning technique was chosen in the pruning phase while the confidence factor was set to 0.25, and the minimum number of samples per leaf was set to 2. Cross-validation (10 folds) is used as an evaluation technique with the three classifiers.

The classification accuracy's average over 10 independent runs was chosen to assess the chosen methods performance. The experiment was implemented on 8 GB of Ram machine with 2.5 GHz Intel Core-i5 CPU, by using Windows 8.1 Pro 64-bit as a platform and using java version "9.0.1".

The following subsections describe the experiment datasets, the parameter settings, and the details of experimental results respectively.

Datasets

Five well-known microarray datasets with different cancer types are used in the experiment. Leukemia and colon microarray datasets can be downloaded from Universidad Pablode Olavide - Bioinformatics Research Group (Anon n.d.). The other datasets, Lung Cancer, Prostate Cancer,

and SRBCT can be downloaded from Vanderbilt University (A. Statnikov 2005). Prostate Tumor, Leukemia, and Colon datasets represent binary classification problems related to the problem of binary cancer detection. On the other hand, Lung Cancer and SRBCT are datasets for multi-class classification that related to the problem of classifying the tumors into different types. Table 1 describes briefly these datasets.

Table 1; The chosen datasets Properties

Dataset Name	Number of Classes	Number of Genes	Number Of Patterns
Colon	2	2000	62
SRBCT	4	2308	83
Leukemia	2	7129	72
Prostate Tumor	2	10509	102
Lung Cancer	5	12600	203

Parameter settings

Some parameters need to be initialized in the proposed method DOC-FS. Table 2 lists the number of genes selected for each dataset after the end of the first phase. The data in Table 2 are obtained through an empirical study. The number of dragonflies equals 100 while, the maximum iterations number equals 500. The rest of the BDF's parameters (*a, s, c, e, d*, and *w*) are adaptively tuned through optimization so that the algorithm can transit from exploration phase to exploitation phase in order to converge.

Experimental Results

The proposed method's performance has been assessed through various datasets using different types of classifies. A comparison between DOC-FS and other methods has been carried out. The result of this comparison has been shown over all of the datasets in Table 3, Table 4, Table 5, Table 6, Table 7, and Table 8. Throughout the paper, the value that represents the highest accuracy is formatted in bold and underline, the value that occupies the second rank is formatted only bold.

Table 3 shows that DOC-FS gains the highest accuracy percentage of classification compared to the other methods over all the datasets. MRMR method shares the first place in classification accuracy percentage with DOC-FS for SRBCT dataset. It also succeeds to get the second-highest classification accuracy for the colon, prostate, and Lung datasets. UFSACO gets the second-highest rating accuracy for the SRBCT dataset. MGSACO obtains the second-highest classification accuracy for the Leukemia dataset.

The result of Table 4 which contains descriptive statistics about Table 3 reveals that DOC-FS is superior to all the other methods with regard to the mean of the average classification accuracy for all datasets with lower standard deviation. Consequently, DOC-FS obtains the highest average classification accuracy compared to other method using the SVM classifier.

It can be observed from Table 5 that DOC-FS beats the other methods with regard to average classification accuracy on all the datasets. The second highest accuracy of classification for the SRBCT, Prostate, and Lung dataset is gotten by the MRMR method. MGSACO method succeeds to get the second highest accuracy of classification for the colon and Leukemia datasets.

Table 2; The number of genes selected for each dataset by the first phase

Dataset	Number of genes selected
Colon	120
SRBCT	450
Leukemia	50
Prostate	350
Lung	50

Table 6 which contains descriptive statistics about Table 5 shows that DOC-FS outperforms the other methods with regard to the mean of average accuracy of classification over all datasets with lower standard deviation. Therefore, DOC-FS gets the highest average accuracy of classification over all the datasets using the NB.

The results of Table 7 show that the average accuracy of classification of DT classifier based on DOC-FS is much better than the other methods for most of the datasets. DOC-FS gets the highest classification accuracy for the Leukemia, Colon, and Prostate datasets. Meanwhile, MRMR gets the highest accuracy of classification for the SRBCT, and Lung datasets. DOC-FS method succeeds to get the second highest accuracy of classification for the SRBCT and Lung datasets. The second highest accuracy of classification for Colon dataset is gotten by the SSF method. Both RRFS and TV obtain the second-highest classification accuracy for the Leukemia dataset. MRMR method gets the second-highest classification accuracy for the Prostate dataset.

It can be observed from Table 8 which contains

descriptive statistics about Table 7 that DOC-FS outperforms the other methods with regard to the mean of the average classification accuracy overall datasets with a lower standard deviation. So, in the case of using the DT classifiers, it's clear that DOC-FS is the best choice.

The conclusion from Table 3, Table 4, Table 5, Table 6, Table 7, and Table 8 that DOC-FS is the best one among the chosen methods with regard to the average accuracy of classification for each of the three classifiers over the five datasets.

The proposed method's execution time has been measured using all of the five datasets. The mean of execution time (in seconds) has been displayed in Figure 5. It's clear that the execution times of DOC-FS increases with the increment of the number of genes selected–in the first phase of DOC-FS - which displayed in Figure 5.

In Figure 5, the x-axis indicates the dataset name while the y-axis indicates the average execution time in seconds.

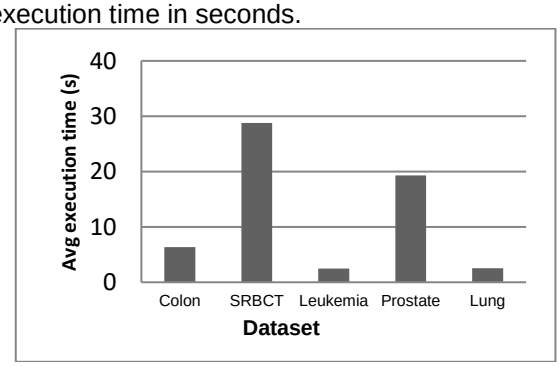

Figure 5; Average execution time in seconds over 10 independent runs

Statistical analysis

Using the non-parametric tests to analyze results obtained by evolutionary algorithms is encouraged by (Derrac et al. 2011; García et al. 2009). In addition, many other feature selection methods used non-parametric tests also (Bermejo et al. 2012; Emary et al. 2015; Tabakhi et al. 2015). For so, Wilcoxon test has been chosen to be performed on the results to demonstrate that the results of the experiment are statistically significant. Wilcoxon test is a nonparametric test which puts ranks to all the scores taken into account as one group. Then, it calculates the total of the ranks of each group (Wilcoxon 1945). Since the Wilcoxon test is a matched-pairs test the proposed method is compared with each method for a certain classifier in a single test as shown in Table 9. During the test, the confidence level of $\alpha=0.05$ has been applied.

In table 9, the symbol "+" means rejecting the null hypothesis, while the first method surpasses the second method. The symbol "–" implies rejecting the null hypothesis, while the first method is beaten by the second method. The symbol "=" means accepting the null hypothesis while the first and the second methods have the same performance. It's clear from Table 9 that all the experimental results are statistically significant except the results of two classifiers in the MRMR method. Our proposed method has higher accuracy but the differences are not significant in the case of SVM and DT. So, in general, we can say that DOC-FS is better than the other methods mentioned in our study.

Table 3; Average classification accuracy of datasets over 10 independent runs using SVM

Datasets	Avg no. of selected genes	Classification accuracy (%)									
		DOC-FS	MRMR	MGSACO	UFSACO	RSM	MC	RRFS	TV	LS	SSF
Colon	18	**86.61**	**83.87**	78.19	78.19	75.46	61.82	75.46	78.19	66.37	74.19
SRBCT	95	**100**	**100**	97.93	**99.31**	91.04	90.35	97.93	96.55	93.11	92.77
Leukemia	22	**97.22**	80.56	**82.06**	58.98	62.36	61.77	76.48	79.42	64.71	80.56
Prostate	20	**93.04**	**84.31**	73.15	59.43	77.15	65.72	69.15	72.00	52.00	76.47
Lung	21	**94.83**	**94.09**	85.72	82.86	64.29	71.43	80.86	72.29	82.00	82.76

Table 4; Descriptive Statistics of SVM classifier

Method	Mean	Std. Deviation	Minimum	Maximum
DOC_FS	**94.34**	**5.05**	**86.61**	**100**
MRMR	**88.566**	**8.14**	**80.56**	**100**
MGSACO	83.41	9.36	73.15	97.93
UFSACO	75.754	17.02	58.98	99.31
RSM	74.06	11.53	62.36	91.04
MC	70.218	11.92	61.77	90.35
RRFS	79.976	10.87	69.15	97.93
TV	79.69	10.00	72	96.55
LS	71.638	16.04	52	93.11
SSF	81.35	7.21	74.19	92.77

Table 5; Average classification accuracy of datasets over 10 independent runs using NB

Datasets	Avg no. of selected genes	Classification accuracy (%)									
		DOC-FS	MRMR	MGSACO	UFSACO	RSM	MC	RRFS	TV	LS	SSF
Colon	18	**84.52**	69.35	**80.00**	71.82	73.64	68.19	67.28	58.19	52.73	69.35
SRBCT	95	**100.00**	98.80	**94.48**	86.90	77.94	83.45	77.25	82.76	75.87	87.95
Leukemia	22	**96.11**	73.61	**92.31**	58.98	57.65	70.59	64.71	67.65	91.18	76.39
Prostate	20	**93.04**	75.49	62.86	60.58	**69.72**	66.29	68.58	66.86	67.43	65.69
Lung	21	**93.99**	93.10	80.00	64.29	76.43	40.96	78.29	68.01	70.01	**80.79**

Table 6; Descriptive Statistics of NB classifier

Method	Mean	Std. Deviation	Minimum	Maximum
DOC_FS	**93.532**	**5.70**	**84.52**	**100**
MRMR	**82.07**	**13.02**	**69.35**	**98.8**
MGSACO	81.93	12.61	62.86	94.48
UFSACO	68.514	11.40	58.98	86.9
RSM	71.076	8.13	57.65	77.94
MC	65.896	15.47	40.96	83.45
RRFS	71.222	6.14	64.71	78.29
TV	68.694	8.84	58.19	82.76
LS	71.444	13.94	52.73	91.18
SSF	76.034	8.89	65.69	87.95

Table 7; Average classification accuracy of datasets over 10 independent runs using DT

Datasets	Avg no. of selected genes	Classification accuracy (%)									
		DOC-FS	MRMR	MGSACO	UFSACO	RSM	MC	RRFS	TV	LS	SSF
Colon	18	**85.32**	70.97	76.37	75.46	71.82	66.37	65.46	68.19	60.91	**83.87**
SRBCT	95	**85.18**	**85.54**	84.14	77.25	68.28	61.38	76.56	68.97	71.04	78.31
Leukemia	22	**85.14**	65.28	76.93	69.24	61.18	67.65	**79.42**	**79.42**	70.59	63.89
Prostate	20	**83.63**	**80.39**	70.29	66.29	66.29	64.00	62.29	61.15	56.01	66.67
Lung	21	**87.64**	**87.68**	80.00	71.43	69.29	68.58	79.72	75.72	78.57	70.94

Table 8; Descriptive Statistics of DT classifier

Method	Mean	Std. Deviation	Minimum	Maximum
DOC_FS	**85.382**	**1.43**	**83.63**	**87.64**
MRMR	**77.972**	**9.58**	**65.28**	**87.68**
MGSACO	77.546	5.09	70.29	84.14
UFSACO	71.934	4.47	66.29	77.25
RSM	67.372	3.99	61.18	71.82
MC	65.596	2.91	61.38	68.58
RRFS	72.69	8.21	62.29	79.72
TV	70.69	7.10	61.15	79.42
LS	67.424	8.94	56.01	78.57
SSF	72.736	8.26	63.89	83.87

Table 9; Results of Wilcoxon's test for the proposed method against the other considered methods through five datasets depending on the accuracy of the three classifiers

	SVM		NB		DT	
Method	**p-value**	**α = 0.05**	**p-value**	**α = 0.05**		**α = 0.05**
MGSACO	0.043	+	0.043	+	0.043	+
UFSACO	0.043	+	0.043	+	0.043	+
RSM	0.043	+	0.043	+	0.043	+
MC	0.043	+	0.043	+	0.043	+
RRFS	0.043	+	0.043	+	0.043	+
TV	0.043	+	0.043	+	0.043	+
LS	0.043	+	0.043	+	0.043	+
SSF	0.043	+	0.043	+	0.043	+
MRMR	0.068	=	0.043	+	0.225	=

DISCUSSION

The main problem in the microarray datasets is the high dimensionality. Most of the genes are considered redundant or irrelevant. Good feature selection methods for microarray datasets eliminate the redundant and irrelevant genes. Univariate feature selection methods such as TV and LS are only able to eliminate irrelevant genes. Furthermore, sample-based feature selection methods like LS can't perform well over microarray datasets because of the little number of samples. However, the multivariate feature selection methods such as MGSACO, UFSACO, RSM, MC, RRFS, and SSF can handle both redundant and irrelevant genes. Therefore, we selected a multivariate feature selection method to be embedded in our proposal.

The proposed method uses the features' predictive performances and intercorrelations in order to lead the search strategy toward a good subset of genes. So, it performs better than univariate feature selection methods like TV and LS. Furthermore, DOC-FS is an SI-based feature selection method which concurrently explores the search space by using many dragonflies in an iterative improvement procedure. Thus, it reaches better results than RSM, MC, RRFS¸ and SSF methods. Furthermore, DOC-FS is based on the dragonfly optimization algorithm that benefits from high exploration convergence of the dragonflies towards good solutions. Using dragonfly as a search strategy conducts to higher performance than MGSACO and UFSACO. DOC-FS accumulative performance is better than the MRMR because DOC-FS is based on population-based mechanism, iterative improvement process, and greedy and stochastic natures which boost the efficiency of DOC-FS compared to the MRMR method.

DOC-FS succeeds to get a high classification accuracy for most of the chosen datasets using relatively a small number of features. Consequently, it'll decrease the computational burden occurred during the classification phase from irrelevant genes. It'll also help in simplifying the gene expression tests to include only a very smaller genes number instead of thousands of genes. Consequently, the testing of cancer cost was minimized significantly. Finally, further biological analysis on the potential biological connection between this little number of genes and development and treatment of cancer will be

possible.

Although the proposed method succeeds to get a high classification accuracy for most of the chosen datasets, it lacks to biological interpretability. In order to improve the interpretability of the genes selected, DOC-FS needs to be integrated with biological knowledge.

CONCLUSION

A new filter feature selection method based on the DF algorithm called DOC-FS has been proposed in this paper to tackle the problem of gene selection in the microarray datasets. The good performance of the DF search strategy and the computational efficiency of the filter approach were merged together in order to boost DOC-FS performance.

DOC-FS performance was evaluated on five microarray datasets by using three classifiers which are support vector machine, decision tree, and naïve Bayes. A comparison between DOC-FS and the most used feature selection methods is done. These methods are MGSACO and UFSACO which are based on ACO, random subspace method (RSM), relevance-redundancy feature selection (RRFS), Simplified silhouette filter (SSF), term variance (TV), mutual correlation (MC), and Laplacian score (LS). Another comparison between the proposed method and the recognized and widely used filter-based features selection method which is the minimal-redundancy- maximal-relevance (MRMR) method has been accomplished. From the experimental results, it can be concluded that the proposed method can select a subset of genes that is of maximum relevance to the decision while having minimal redundancy between themselves. Furthermore, experimental results show that the classification accuracy of the proposed method is superior to that of the other filter-based feature selection methods for different datasets over all three classifiers. It can be confirmed that over different classifiers DOC-FS has good results. Finally, it can be verified that DOC-FS can be used with many classifiers to obtain a fast automatic diagnostic system.

CONFLICT OF INTEREST

The authors declared that present study was performed in absence of any conflict of interest.

ACKNOWLEGEMENT

The author would like to thank the staff of microbiology department in College of Science, Helwan University, Egypt, for their great biological support.

AUTHOR CONTRIBUTIONS

MG designed the proposed algorithm and performed the experiments and also wrote the manusscript. EN participated in designing the proposed algorithm and performed the statistical analysis. AB and EN reviewd the manuscript. SF and AB designed the research project.

REFERENCES

A. Statnikov, C.F.A.I.T., 2005. Gene Expression Model Selector. Available at: http://www.gems-system.org.

Aghdam, M.H., Ghasem-Aghaee, N. & Basiri, M.E., 2009. Text feature selection using ant colony optimization. *Expert systems with applications*, 36(3), pp.6843–6853.

Anon, Dataset Repository, Bioinformatics Research Group. Available at: ⟨http://www.upo.es/eps/bigs/datasets.html⟩ [Accessed 2014].

Bermejo, P. et al., 2012. Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking. *Knowledge-Based Systems*, 25(1), pp.35–44.

Bertoni, A., Folgieri, R. & Valentini, G., 2005. Bio-molecular cancer prediction with random subspace ensembles of support vector machines. *Neurocomputing*, 63, pp.535–539.

Bolón-Canedo, V. et al., 2014. A review of microarray datasets and applied feature selection methods. *Information Sciences*, 282, pp.111–135.

Cai, R. et al., 2009. An efficient gene selection algorithm based on mutual information. *Neurocomputing*, 72(4-6), pp.991–999.

Covões, T.F. & Hruschka, E.R., 2011. Towards improving cluster-based feature selection with a simplified silhouette filter. *Information Sciences*, 181(18), pp.3766–3782.

Dashtban, M. & Balafar, M., 2017. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. *Genomics*, 109(2), pp.91–107.

Derrac, J. et al., 2011. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1), pp.3–18.

Ding, C. & Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02), pp.185–205.

Dorigo, M. & Stützle, T., 2003. The ant colony optimization metaheuristic: Algorithms, applications, and advances. In *Handbook of metaheuristics*. Springer, pp. 250–285.

Elyasigomari, V. et al., 2017. Development of a two-stage gene selection method that incorporates a novel hybrid approach using the cuckoo optimization algorithm and harmony search for cancer classification. *Journal of biomedical informatics*, 67, pp.11–20.

Emary, E. et al., 2015. Firefly Optimization Algorithm for Feature Selection. In *Proceedings of the 7th Balkan Conference on Informatics Conference*. p. 26.

Ferreira, A.J. & Figueiredo, M.A., 2012a. An unsupervised approach to feature discretization and selection. *Pattern Recognition*, 45(9), pp.3048–3060.

Ferreira, A.J. & Figueiredo, M.A., 2012b. Efficient feature selection filters for high-dimensional data. *Pattern Recognition Letters*, 33(13), pp.1794–1804.

García, S. et al., 2009. A study on the use of non-parametric tests for analyzing the evolutionary algorithms' behaviour: a case study on the CEC'2005 special session on real parameter optimization. *Journal of Heuristics*, 15(6), p.617.

Ghazavi, S.N. & Liao, T.W., 2008. Medical data mining by fuzzy modeling with selected features. *Artificial Intelligence in Medicine*, 43(3), pp.195–206.

Gheyas, I.A. & Smith, L.S., 2010. Feature subset selection in large dimensionality domains. *Pattern recognition*, 43(1), pp.5–13.

Golub, T.R. et al., 1999. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439), pp.531–537.

Gutiérrez-Avilés, D. et al., 2014. TriGen: A genetic algorithm to mine triclusters in temporal gene expression data. *Neurocomputing*, 132, pp.42–53.

Guyon, I. et al., 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3), pp.389–422.

Haindl, M. et al., 2006. Feature selection based on mutual correlation. In *Iberoamerican Congress on Pattern Recognition*. pp. 569–577.

Hall, M. et al., 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1), pp.10–18.

Hall, M.A., 1999. *Correlation-based feature selection for machine learning*.

Hall, M.A., 1999. Feature selection for discrete and numeric class machine learning.

He, X., Cai, D. & Niyogi, P., 2006. Laplacian score for feature selection. *Advances in Neural Information Processing Systems*, 18, p.507.

Inza, I. et al., 2002. Gene selection by sequential search wrapper approaches in microarray cancer class prediction. *Journal of Intelligent \& Fuzzy Systems*, 12(1), pp.25–33.

Inza, I. et al., 2004. Filter versus wrapper gene selection approaches in DNA microarray domains. *Artificial intelligence in medicine*, 31(2), pp.91–103.

Kabir, M.M., Shahjahan, M. & Murase, K., 2012. A new hybrid ant colony optimization algorithm for feature selection. *Expert Systems with Applications*, 39(3), pp.3747–3763.

Kanan, H.R. & Faez, K., 2008. An improved feature selection method based on ant colony optimization (ACO) evaluated on face recognition system. *Applied Mathematics and Computation*, 205(2), pp.716–725.

Lai, C., Reinders, M.J. & Wessels, L., 2006. Random subspace method for multivariate feature selection. *Pattern recognition letters*, 27(10), pp.1067–1076.

Lai, C., Reinders, M.J., van't Veer, L.J., et al., 2006. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC bioinformatics*, 7(1), p.235.

Lazar, C. et al., 2012. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4), pp.1106–1119.

Lee, C.-P. & Leu, Y., 2011. A novel hybrid feature selection method for microarray data analysis. *Applied Soft Computing*, 11(1),

pp.208–213.

Leung, Y. & Hung, Y., 2010. A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 7(1), pp.108–117.

Li, X. & Zhao, H., 2009. Weighted random subspace method for high dimensional data classification. *Statistics and its Interface*, 2(2), p.153.

Li, Y. et al., 2013. An ant colony optimization based dimension reduction method for high-dimensional datasets. *Journal of Bionic Engineering*, 10(2), pp.231–241.

Liao, B. et al., 2014. Gene selection using locality sensitive Laplacian score. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 11(6), pp.1146–1156.

Liu, H. & Yu, L., 2005. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on knowledge and data engineering*, 17(4), pp.491–502.

Lu, X. et al., 2014. A novel feature selection method based on correlation-based feature selection in cancer recognition. *Journal of computational and Theoretical nanoscience*, 11(2), pp.427–433.

Mafarja, M.M. et al., 2017. Binary dragonfly algorithm for feature selection. In *New Trends in Computing Sciences (ICTCS), 2017 International Conference on*. pp. 12–17.

Marinakis, Y. et al., 2009. Ant colony and particle swarm optimization for financial classification problems. *Expert Systems with Applications*, 36(7), pp.10604–10611.

Martinez, E., Alvarez, M.M. & Trevino, V., 2010. Compact cancer biomarkers discovery using a swarm intelligence feature selection algorithm. *Computational biology and chemistry*, 34(4), pp.244–250.

Medjahed, S.A. et al., 2016. Kernel-Based Learning and Feature Selection Analysis for Cancer Diagnosis. *Applied Soft Computing*.

Mirjalili, S. & Lewis, A., 2013. S-shaped versus V-shaped transfer functions for binary particle swarm optimization. *Swarm and Evolutionary Computation*, 9, pp.1–14.

Mirjalili, S., 2016. Dragonfly algorithm: a new meta-heuristic optimization technique for solving single-objective, discrete, and multi-objective problems. *Neural Computing and Applications*, 27(4), pp.1053–1073.

Nguyen, T. et al., 2015. Hidden Markov models for cancer classification using gene expression profiles. *Information Sciences*, 316, pp.293–307.

Niijima, S. & Okuno, Y., 2009. Laplacian linear discriminant analysis approach to unsupervised feature selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(4), pp.605–614.

Peng, H., Long, F. & Ding, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8), pp.1226–1238.

Quinlan, J.R., 1986. Induction of decision trees. *Machine learning*, 1(1), pp.81–106.

Raileanu, L.E. & Stoffel, K., 2004. Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, 41(1), pp.77–93.

Ramón & De Andres, S.A., 2006. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7(1), p.3.

Rapaport, F. et al., 2007. Classification of microarray data using gene networks. *BMC bioinformatics*, 8(1), p.35.

Rubio-Escudero, C. et al., 2008. Classification of gene expression profiles: comparison of K-means and expectation maximization algorithms. In *2008 Eighth International Conference on Hybrid Intelligent Systems*. pp. 831–836.

Saeys, Y., Inza, I. & Larrañaga, P., 2007. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19), pp.2507–2517.

Sahu, B. & Mishra, D., 2012. A novel feature selection algorithm using particle swarm optimization for cancer microarray data. *Procedia Engineering*, 38, pp.27–31.

Salem, H., Attiya, G. & El-Fishawy, N., 2017. Classification of human cancer diseases by gene expression profiles. *Applied Soft Computing*, 50, pp.124–134.

Sergoios Theodoridis, K.K., 2008. *Pattern Recognition (Fourth Edition)*, Academic Press, Oxford.

Shi, Y. & others, 2001. Particle swarm optimization: developments, applications and resources. In *evolutionary computation, 2001. Proceedings of the 2001 Congress on*. pp. 81–86.

Shyamsundar, R. et al., 2005. A DNA microarray survey of gene expression in normal human

tissues. *Genome biology*, 6(3), p.R22.

Srivastava, A. et al., 2013. Hybrid firefly based simultaneous gene selection and cancer classification using support vector machines and random forests. In *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*. pp. 485–494.

Tabakhi, S. et al., 2015. Gene selection for microarray data classification using a novel ant colony optimization. *Neurocomputing*.

Tabakhi, S., Moradi, P. & Akhlaghian, F., 2014. An unsupervised feature selection algorithm based on ant colony optimization. *Engineering Applications of Artificial Intelligence*, 32, pp.112–123.

Umamaheswari, K. & Dhivya, M., 2016. D-MBPSO: An Unsupervised Feature Selection Algorithm Based on PSO. In *Innovations in Bio-Inspired Computing and Applications*. Springer, pp. 359–369.

Wang, G. et al., 2013. Selecting feature subset for high dimensional data via the propositional FOIL rules. *Pattern Recognition*, 46(1), pp.199–214.

Wilcoxon, F., 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6), p.80.

Witten, I.H. & Frank, E., 2005. *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann.

Yu, H. et al., 2009. A modified ant colony optimization algorithm for tumor marker gene selection. *Genomics, proteomics \& bioinformatics*, 7(4), pp.200–208.

Yu, L. & Liu, H., 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*. pp. 856–863.

Yu, L. & Liu, H., 2004. Efficient feature selection via analysis of relevance and redundancy. *Journal of machine learning research*, 5(Oct), pp.1205–1224.

Zhao, W. et al., 2011. A novel framework for gene selection. *Int J Adv Comput Technol*, 3(3), pp.184–91.

Zibakhsh, A. & Abadeh, M.S., 2013. Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function. *Engineering Applications of Artificial Intelligence*, 26(4), pp.1274–1281.