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(2): 1323-1336. **OPEN ACCESS**

A new hybrid approach for feature selection and predicting of protein interaction network in lung cancer

Zeinab Abd El Haliem[1], Mohammad Nassef[2], Amr Badr[2] and Khaled T. Wassif[2]

[1]**Department of Computer Science, Faculty of Computer Science, Modern Sciences and Arts University, Giza, Egypt.**
[2]**Department of Computer Science, Faculty of Computer and Information, Cairo University, Giza, Egypt.**
.
***Correspondence:** ztaha@msa.eun.eg **Accepted: 12**Apr. 2019 **Published online:09** Mar. 2019

Different computational and evolutionary methods have been employed in the last decade for selecting important molecular features from biological data. Extracting information from microarray data is extremely important and complex task due to the high dimensionality of its datasets. Feature selection is a very important aspect of the analysis that helps in identifying the important genes that can be used in a further biological analysis. This paper proposes a new hybridization between the Flower Pollination and Differential Evolution algorithms for optimizing feature selection parameters and to find out the most important subset of features over gene expression profiles of lung cancer. The results showed that the hybrid approach has a better capability in searching for the best solutions compared to applying each algorithm independently. SLC5A1 gene was identified as a biomarker gene of lung cancer. By constructing the protein-protein interaction network for the extracted genes, a direct interaction has been detected between the SLC5A1 and EGFR genes, where the latter is known to have an important role in the mutation process of lung cells.

Keywords: Evolutionary algorithms; Flower pollination algorithm; Differential evolution; Gene expression; Protein-Protein Interaction

INTRODUCTION

Cancer happens when some cells in the body reproduce immensely leading to out-of-control reproduction. Cancer arises from the mutation/alteration of one or more normal genes in the cell that are called *oncogenes*. Cells that are old or not functioning properly normally self-destruct and are replaced by new cells. However, the cancerous cells go rapidly through reproduction of millions of newly cancerous cells.

Lung cancer is the second kind of cancer causing death worldwide. It is known to be the leading cause of cancer death in both men and women in the United States (John et al., 1994). The everyday deaths caused by lung cancer are greater than the deaths caused by other types of cancer like breast, colon, and prostate cancers combined. Based on statistics by the American Cancer Society, it is believed that there are 220,000 new lung cancer cases per year; death per year is about 160,000 and 5-year survival rate of all stages is 15% (Kancherla and Mukkamala 2012). However the 5-year survival rate of localized stage is about 50%. Cigarette smoking is the number one cause of lung cancer. Lung cancer moreover can be caused by using other types of tobacco, breathing secondhand smoke, being exposed to substances and having a family history of lung cancer.

Feature selection plays an important role in knowledge discovery and in the field of data mining as many problems need to be solved by selecting a subset of discriminative features. It is also used to improve the accuracy of classification and to reduce the computational time of algorithms. There are two approaches for feature selection methods; Filter-based approach which is not directly related to the classification performance and the evaluation of features is based on the general characteristics of the data without any consideration to any mining algorithms. In contrast, the Wrapper-based approach is related to the classification performance without redundancy in the selected features, therefore it requires a mining algorithm and uses the performance to determine the best features from the feature sets (Sayed et al., 2016). Generally, feature selection is used to minimize the feature space and improve classification accuracy to get the best-optimized solution. In biological systems it helps in identifying the important genes that will reflect on the prediction systems.

Predicting protein- protein interaction networks is a very important and necessary operation because it provides a global picture of cellular functions and biological processes. The level of interaction between some protein and others depends on the nature and functionality of that protein. Although some proteins highly interact with others, other proteins have fewer interactions. The dysfunction of some interactions can cause many serious diseases including cancer. To solve the cancer classification problem, the molecular mechanisms of diseases through human interaction networks should be understood. The discovery of gene features has a great impact on distinguishing between normal and tumor cancer samples, understanding the molecular mechanisms and systems, and exploring new ways for treatments. The STRING database can be used to search about possible protein-protein Interactions (PPI) to understand more about the genomic data and the relation between different proteins. There are variety of computational methods that are used for predicting and detecting possible biomarkers for cancer from microarray data such as in (Guyon et al., 2002). Furthermore, the field of evolutionary algorithms is of a great interest for many researchers and developers, such algorithms are usually used for optimizing feature selection and other computational methods, In addition, they have been used for discovering molecular features in biological data (Coello et al.,2015). Evolutionary algorithms take the advantage of simplicity and robustness; some of these algorithms have been developed recently to solve biological optimization problems by emulating the behavior of bees, ants, and fireflies atc. As in (Abraham et al., 2008, Dorigo et al., 2006 and Karaboga et al., 2007).

The Flower Pollination Algorithm (FPA) is a new meta-heuristic computational technique proposed by Xin-She Yang (2012); it is one of the recent algorithms which use the behavior of the pollination process of flowers to find optimal feature sets. Many research efforts have been proposed to solve feature selection problem using flower pollination algorithm (Yang et al., 2013, Abdel-Raouf et al., 2014, and Rodrigues et al., 2015). And a successful hybridization with binary clonal selection algorithms has been proposed by (Sayed et al., 2016).

The Differential Evolution Algorithm (DEA) is initially introduced by (Storn and Price 1997). DEA emerged as the best competitive and effective optimizer amongst all the evolutionary algorithms in self-adaptive problems and real-valued functions (Qin and Suganthan 2005). Several previous studies were conducted using DEA for solving feature selection as in (He et al., 2009, Sikdar et al., 2012), (Das and Suganthan 2011) and (Cai and Du 2014). Consequently, DEA varies in many application domains such as mechanical engineering, industrial communication, and pattern recognition algorithms. Successful hybridization of differential evolution with biogeography-based optimization were used for global numerical optimization in

(Gong et al., 2010) with practical swarm optimization for feature selection (Robic and Filipic 2005), and multi objective optimization with differential evolution as in (Zhang and Xie 2003).

This paper proposes a hybrid approach which uses FPA and DEA for feature selection of lung cancer gene expression data and constructing the protein-protein interaction network for the proposed genes/proteins. The Optimum Path Forest (OPF) classifier is used and its accuracy was used also as a fitness function to be maximized. OPF classifier was introduced by (Papa et al., 2009). OPF has the advantage of accelerating capability for training data without any changes in the accuracy level compared to other classifiers (Papa et al., 2012) and (Nakamura et al., 2012). It was rumored to own faster-training capabilities from 10 up to thousand times rather than the other classifiers without affecting the accuracy (Papa et al., 2009). OPF has been demonstrated to be as effective as Support Vector Machine Classifier but OPF is faster in the training. The performance of the proposed algorithm is tested and evaluated using comparable results from using each algorithm individually in The Cancer Genomic Atlas (TCGA) gene expression lung cancer data (https://cancergenome.nih.gov/) and another public data set (GSE10072), and notated that the hybrid approach gives better results and higher efficiency.

The rest of this paper is structured as follows: Section 2 describes the fundamentals of feature selection, characteristics of FPA and DEA based on feature selection. Section 3 describes the proposed approach, showing its effectiveness in solving that problem. Section 4 shows the evaluation and results for applying each algorithm individually compared with a hybrid one. Finally, Section 5 concludes the proposed model and highlights some points for future work.

MATERIALS AND METHODS

This section describes the feature selection problem and its fundamentals. Next, the characteristics of the algorithms used in this paper will be presented.

Feature Selection

The feature selection process is a searching process for the best feature(s) in the entire feature set, thus, it is an optimization process. Additionally, (John et al. 1994) defined the relevance of a feature through a probability distribution over the feature values and different labels. Consequently, a feature has a strong relationship to the relevance when the probability distribution is affected whether that feature is eliminated from the feature set or not. However, a feature may become less relevant under a certain combination of features. Hence, the optimum feature subset consists of the relevant features only.

Microarray dataset contains a few samples comparable with thousands of features that are actually involved in these samples. Microarray data consists of a collection of samples S-labeled to a specific class, each sample in S can be represented as a vector $x_i = \{x_{i1}, x_{i2} \ldots x_{in}\}$ with *n* features (genes) in the gene expression profile.

Flower Pollination Algorithm (FPA)

Flower Pollination Algorithm is a new meta-heuristic computational search technique proposed by (Xin-She Yang 2012). Many research efforts have been proposed to solve feature selection problem using flower pollination algorithm (Yang et al., 2013, Abdel-Raouf et al., 2014, and Rodrigues et al., 2015). And a successful hybridization with binary clonal selection algorithms has been proposed by [3]. FPA is inspired by the pollination process of flowering plants. Pollination is the process of relocating pollens from the male to the female stigma of a flower (Sayed et al., 2016). The goal of all living creatures is to produce offspring for the next generation hoping that offspring will have good genesis better than their parents. Pollination process can be done in two ways; self-pollination or cross-pollination. Cross-pollination occurs when the pollens of one plant are transferred to another flower from another plant. Biotic creatures like birds, insects, bees, etc. are examples of pollinators of cross-pollination process. In contrast, self-pollination occurs when the flower makes pollination in the same plant, such as peach flowers in which pollen of the same flower or different flowers of the same plant, and this process is done when there is no reliable pollinator is available.

In addition, bees and birds may behave as Levy flight behavior (Sayed et al., 2016); (Yang et al., 2012) with a jump or fly distance steps that obey a levy distribution. Moreover, due to physical factors such as wind, local pollination can have an increment step or fraction in the overall pollination actions. But in the global pollination step, the flower pollens are carried by pollinators such as insects, and pollens can travel over longer distances.

(Yang 2012) stated that FPA follows four basic rules listed as follows:

Abiotic method and self-pollination are considered as Local Pollination.

Biotic method and cross-pollination are considered as Global Pollination.

Local and Global Pollination is restricted by a switching probability p ∊ [0, 1]. Due to physical issues such as wind, the local pollination has a significant probability p in the whole pollination activities.

Duplication probability is relative to the similarity of two flowers involved and regarding that the flower consistency is considered.

Global pollination can be expressed by the following equation which is a mathematical representation of Rules (2) and (4):

$$x_i^{(t+1)} = x_i^t + \alpha\, L(\lambda)(g^* - x_i^t) \qquad \textbf{(1)}$$

Where

$$L(\lambda) = \frac{\lambda.\Gamma(\lambda) . \sin(\lambda)}{\pi} . \frac{1}{s^{1+\lambda}} \quad , s > 0 \qquad \textbf{(2)}$$

Where x_i^t is the pollen i (solution list) at the iteration t and g* is the current best solution found among all solutions at the current generation, while α is the scaling factor to control the step sizes, L(λ) is the Lévy flight step size corresponding to the strength of the pollination and Γ(λ) stands for the gamma function, where the value of λ is in the range of 1 ≤λ ≤2.While insects can move over a long distance with many steps to reach, therefore Lévy flights can be used to handle this issue well. At this step Local pollination in Rules (1) and (3) can be represented mathematically by the following equation:

$$x_i^{(t+1)} = x_i^t + \varepsilon(x_j^t - x_k^t) \qquad \textbf{(3)}$$

Where $x_j^t$and x_k^tstand for the pollen from different flowers j and k of the same plant class. The switching probability P is used in (Rule 3) to mimic the local and the global flower pollination.

Finally, the main objective of the flower pollination algorithm is achieving and protecting the best and the optimal offspring of plants. Algorithm (1) represents the pseudo-code for the basic flower pollination algorithm (FPA).

Algorithm 1 Flower Pollination Algorithm

```
Objective min or max $f(x)$, $x = (x_1 \ldots x_n)$;
Initialize a population of $n$ flowers/pollen gametes with random solutions;
Find the best solution $g^*$ in the initial population;
Define a switch probability $p \in [0, 1]$;
While ($t$ <MaxGeneration) do
    For i = 1: n (all $n$ flowers in the population) do
        If $rand < p$, then
            Draw a ($d$-dimensional) step vector $L$ which obeys a Lévy flight distribution; and
            Global pollination;
        Else
            Draw € from a uniform distribution in p [0, 1];
            Randomly choose $j$ and $k$ among all the solutions;
            Do local pollination;
        End if
        Evaluate new solutions;
        if new solutions are better, update them in the population;
    End for
    Find the current best solution $g^*$;
End while
```

Differential Evolution Algorithm (DEA)

A differential evolution algorithm is a population-based algorithm (Nakamura et al., 2012); it was proposed by Storn and Price after some adaptation and handling issues in the old version by (Storn and Price 1996). DEA was used for solving many problems like a feature selection problem as in (He et al., 2009, Sikdar et al., 2012), (Das and Suganthan 2011) and (Cai and Du 2014) by proposing a self-adaptive DE (SaDE) algorithm, in which both trial vector generation strategies and their associated control parameter values are gradually self-adapted by learning from their previous experiences in generating promising solutions. Consequently, DEA is used in many application domains such as mechanical engineering, industrial communication, and pattern recognition algorithms. The Differential Evolution has been successfully hybridized with many optimization techniques including biogeography-based optimization for global numerical optimization (Gong et al., 2010), practical swarm optimization for feature selection (Robic and Filipic 2005), and multi objective optimization (Zhang and Xie 2003).

DEA was considered as a simple and a fast technique for getting the best solution; it emerged as the best optimizer among all the evolutionary algorithms in self adaptive problems and real-valued functions. DEA has an advantage of having a variation on the mutation scheme of the algorithm. Unlike genetic algorithm or any other evolutionary algorithm, it performs mutation before crossover to generate offspring that is used within the crossover process. Moreover, the step sizes of the mutation process are not sampled from prior knowledge (Akutekwe et al., 2014). The standard DEA has three main evolutionary operations (Mutation, Crossover, and Selection operation) to reach the optimum solution over the whole search space, as well as the fitness function evaluates the offspring in each iteration and updates offspring in the population. DEA aims at partitioning a population X into a number of patterns NP which can be represented as a feature vector or individuals in D-dimensional vector as shown in equation (4).

$$X_{i,k} = \{ X_{i,k}^1, X_{i,k}^2, \ldots \ldots ., X_{i,k}^D \} \qquad \textbf{(4)}$$

Where i= 1, 2… NP

The initial population should enclose the entire search space by randomizing individuals using the recommended minimum or maximum bounds.

$X_{min} = \{ X_{min}^1, X_{min}^2, \ldots ., X_{min}^D \}$ And

$$X_{max} = \{ X_{max}^1, X_{max}^2, \dots., X_{max}^D \}$$

Equation (5) represents the initialization of the *j*^th^ parameter in the solution *i*^th^ when the generation K=0.

$$X_{i,0}^j = X_{min}^j + rand(0,1).(X_{max}^j - X_{min}^j) \quad \textbf{(5)}$$

Where j= 1, 2… D and rand (0, 1) is a random value between 0 and 1.

After the initialization is done, the next step is the mutation which produces a mutant vector Vi, k for each individual vector Xi, k in the K generation.

$$V_{i,k} = \{V_{i,k}^1, V_{i,k}^2, \dots, V_{i,k}^D\}$$

There are many ways to generate the mutant vector, but DEA usually uses five most frequent strategies to generate the mutant vector which seams as follow:

(1) DE/rand/1:

$$V_{i,k} = X_{r_1^i,k} + F.(X_{r_2^i,k} - X_{r_3^i,k}) \quad \textbf{(6)}$$

(2) DE/best/1:

$$V_{i,k} = X_{best,k} + F.(X_{r_1^i,k} - X_{r_2^i,k}) \quad \textbf{(7)}$$

(3) DE/rand to best/1:

$$V_{i,k} = X_{best,k} + F.\left(X_{best,k} - X_{r_3^i,k}\right) + F.\left(X_{r_1^i,k} - X_{r_2^i,k}\right) \textbf{(8)}$$

(4) DE/best/2:

$$V_{i,k} = X_{best,k} + F.\left(X_{r_1^i,k} - X_{r_2^i,k}\right) + F.\left(X_{r_3^i,k} - X_{r_4^i,k}\right) \textbf{(9)}$$

(5) DE/rand/2:

$$V_{i,k} = X_{r_1^i,k} + F.\left(X_{r_2^i,k} - X_{r_3^i,k}\right) + F.\left(X_{r_4^i,k} - X_{r_5^i,k}\right) \textbf{(10)}$$

Where $\boldsymbol{r_1^i, r_2^i, r_3^i, r_4^i, r_5^i}$ are generated in the range [1, NP], and they are mutually exclusive integers. *F* is a control parameter proposed by Storn and Price which is a constant value between [0, 1].

The next operation is the crossover process which is applied to each target vector in the population. There are two main types of crossover in DEA: *binominal* and *Exponential*. In binominal crossover, each variable x_i is exchanged with conditional offspring x_{off} with probability of crossover rate (CR) which represented as a constant value in the range [0, 1] determined by the user.

$$X_{off^{[j]}} = \begin{cases} X_{i,off^{[j]}} & if\ (\varphi(0,1) \le CR\ OR \\ X_i[j] & otherwise \end{cases} \quad (j = j_{rand})\textbf{(11)}$$

Where φ(0, 1) is a distributed random number and *j* is the index of the gene being selected.

Alternatively, in the exponential crossover, some (a segment of) genes of the parent are copied into the offspring (child) until a sequence of random numbers smaller than the threshold CR is reached (Akutekwe et al., 2014). The last process of DEA is the Selection process which differs from any other evolutionary algorithm. The new generation is selected from the old generation and its corresponding trial vectors. All the processes of Mutation, Crossover and Selection are running continuously until reaching the optimum offspring which represents the best solution at that time. Algorithm (2) represents the pseudo-code for DEA

Algorithm 2: Differential Evolution Algorithm

```
Generate an initial population of $N_p$ individuals and parameter setting of control parameters (F, CR);
Evaluate Fitness Function of each solution
While stopping condition is not me
Do
For each $x_i$ in $N_p$ do
    Generate offspring $x_{off}$ by applying mutation process
    Generate offspring $x_{off}$ by applying crossover process
    Evaluate Fitness Function of offspring $x_{off}$
    Check performance of old & new offspring
End For
For each $x_i$ solution in $N_p$ do
    Select the best solutions between chromosomes (offspring) and do all the reprocessing operations.
End For
End While
Return best solution
```

Optimum Path Forest (OPF)Classifier

Papa et al. presented OPF as a simple, fast, efficient, and parameter independent classifier (Papa et al., 2009, 2012). OPF is a supervised classification method, and the training dataset can be represented as a complete graph. It represents the samples as graph nodes whose arcs are weighted by using any distance function. In the graph, each node is represented as a feature vector, and each edge connects a pair of nodes, constituting a fully connected graph (Sayed et al., 2016).

In our study, the dataset Z of lung cancer is divided into two parts Z_1 and Z_2 where Z_1 is the training set and Z_2 is the testing set, and Z is a fully labeled dataset. Let (Z_1, A) is a complete graph whose nodes are the samples in this set and any pair of samples represents an arc in A=Z_1 X Z_1. Let π_S be a path in the graph in sample S ∈ Z_1 (training set), and (π_S. (S, t)) Is the concatenation between π_S and the arc (s, t) where t∈ Z_1, and S ⊂ Z_1 as a set of key prototypes of all classes (samples). The past cost can be computed by using the following equation:

$$f_{max} = \begin{cases} 0 & if s \in S, \\ +\infty & \text{otherwise}, \end{cases} \quad \textbf{(12)}$$

$$f(\pi_s . (s, t)) = \max\{f(\pi_s), d(s, t)\}, \quad \textbf{(13)}$$

Where *d(s, t)* is the distance between node *s* and node *t.*

A group of prototypes which can be represented as S^* (an optimal set of prototypes) can be found using the representation of Minimum Spanning Tree (MST) in the complete graph (Z_1, A). A MST can be described as optimum when the sum of its arc weights is the lowest amount compared to any other spanning tree in the complete graph. A MST contains just one optimum path tree for any selected root node, and to get it, the closest elements of this tree have to be selected with different labels in Z_1. Every pair of samples in this MST is connected by a single path which can be evaluated as minimum or not by equation (12).

Consequently, in the graph, nodes represent all the samples of Z_1, and the arcs are weighted by the distance *d* between any adjacent samples.

The training phase of this classifier starts with nodes (prototypes) to minimize the cost between each pair or sample in the training set samples. After that, it gets an optimum path forest which can be described as a collection of optimum path trees rooted at each node or prototype. Conversely, in the testing/classification phase all the arcs are taken into consideration especially those connecting a *t* sample in the testing data Z_2 with samples s ∈ Z_1 (training set), so the sample *t* was a part of the training graph. The optimum path P^*(t) can be found by evaluating all possible paths from S^* to the sample *t*, and label *t* with the most strongly connected prototype in all paths S^* by λ(R (t)) ∈ S^*, where λ (t) is the function that assigns the correct class label, and R (t) is the function that gets the root of *t* and this root is one of the prototypes R (t) ∈S. This path can be identified by calculating the optimum cost equation (14) as follows:

$$C(t) = \min\{\max\{C(s), d(s, t)\}\}, \forall\, s \in Z_1 \quad \textbf{(14)}$$

According to Eq. (14), it can be assumed that the node P (t) is the predecessor in the optimum path P*(t), and S*∈ Z1 is the node that satisfies the equation too. Given that L(S*) = λ (R (t)) as the class t.

STRING DATABASE- PROTEIN-PROTEIN INTERACTION

STRING is an online tool and database resource (http://string-db.org) that provides critical interactions and assessment of protein-protein interactions through direct or indirect (physical or functional) associations. STRING database was developed by a consortium of academic institutions, including CPR, EMBL, KU, SIB, TUD and UZH. The latest version 10 in STRING database contains information about around 9.6 million proteins which covers more than 2000 organisms. It provides us with algorithms for transferring interaction information between organisms and hierarchical and self-consistent annotations for all interacting proteins. Furthermore, STRING is used for retrieving the interactions of genes/proteins in a graph using levels of hierarchy by representing and grouping these genes or proteins into families. The STRING database contains a lot of information from different sources, including experimental data, computational prediction methods and public text of data. It is freely accessible and it is regularly updated (Szklarczyk et al., 2011).

The Protein-Protein Interaction (PPI) networks of a specific or multiple proteins shows the interaction of proteins which have the short distances between them and tend to have the same biological functions. PPI networks are very important and critical assets for understanding the level of the system of cellular processes by using it for filtering or accessing functional genomic data and for providing functional and evolutionary properties about proteins.

PROPOSED SOLUTION

In this paper, a new hybrid approach between FPA and DEA is proposed and developed to solve the problem of feature selection in gene expression of lung cancer data profiles. Subsequently searching the database for protein to construct PPI network which describes the interactions between the output proteins with the other proteins in the specified area related to this biological issue.

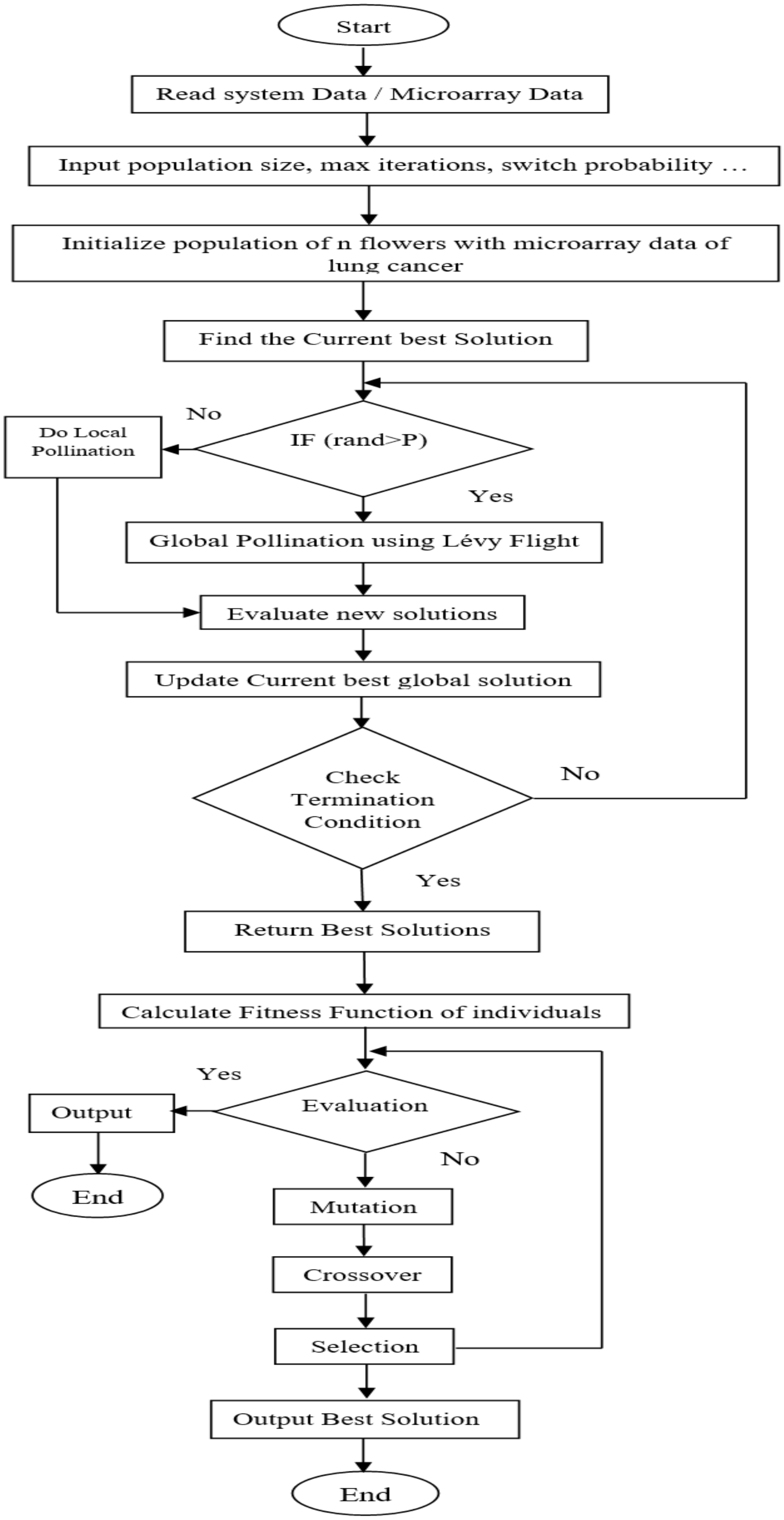
Start
Read system Data / Microarray Data
Input population size, max iterations, switch probability …
Initialize population of n flowers with microarray data of lung cancer
Find the Current best Solution
IF (rand>P)
No
Do Local Pollination
Yes
Global Pollination using Lévy Flight
Evaluate new solutions
Update Current best global solution
Check Termination Condition
No
Yes
Return Best Solutions
Calculate Fitness Function of individuals
Evaluation
Yes
Output
End
No
Mutation
Crossover
Selection
Output Best Solution
End

Figure 1: Flow Chart of Proposed Methodology

Genes with Hybrid Features

In our study, flower pollination and differential evolution algorithms are jointly and cooperatively used to solve the problem of feature selection.

Hybridization makes use of the strengths of each algorithm to form a hybrid algorithm that can be efficiently used in solving specific problems.

Hybridization usually results in some improvements mainly in the accuracy or computational speed of the system. FPA may be shared and used as a sub-algorithm to set the optimal parameters for DEA, whereas different components of DEA like mutation, crossover, and selection are used to improve the result of this hybrid system.

FPA is a very good searching algorithm which has good characteristics through using Lévy flight. DEA also has many advantages for optimizing problems. Hence, the composition of the two algorithms has been used in this study to maximize the benefits through working together.

The flowchart of the proposed hybrid methodology is shown in Figure 1. The figure mainly describes the three important stages, i.e., applying FPA, DEA, and then OPF.

In the First Stage, the FPA is used to select top genes with the highest scores (best genes) through applying Algorithm (1) considering Local and global pollination, and the OPF is trained over the data within each new generation, and its parameters and feature subset are dynamically optimized.

The output genes will then be utilized for the Second Stage as DEA applied Algorithm (2) with the required fitness function. Mutation is used to balance the search space exploration during the search process to get genes with best features.

Finally, search the STRING database about possible PPI to understand more about the genomic data and the relation between these genes/proteins. The discovery of the gene features has a great impact for discovering and distinguishing between normal and tumor cancer, helping in understanding mechanisms and systems, and useful for exploring new ways for treatments.

RESULTS

EXPERIMENTAL RESULTS

This section starts with a brief description of the experimental microarray datasets attempted in this work. After that, the results of the hybrid algorithm are discussed after the illustration of the used evaluation criteria.

Datasets

There are many microarray datasets published from cancer gene expression studies. In this paper, the concentration is in Lung cancer as a specific type of cancer. To assess the effectiveness of the proposed approach two datasets are used in this work; such datasets differ on the number of samples, features and classes. The Cancer Genomic Atlas (TCGA) data portal is used to get the genomic matrix (gene expression) of lung cancer using the Human Infinium 450k assay for 4034 cancer and normal tissue samples and another public data set (GSE10072) was used also to evaluate the proposed hybrid method (https://cancergenome.nih.gov/). Table 1 presents the datasets used in this work.

Evaluation Criteria

The OPF is used as a classifier to evaluate the classification individual performance of the original FPA and DEA algorithms. Following the evaluation experiment used in (Rodrigues et al., 2015). Initially, the dataset, which denoted in the previous section (OPF section) as Z, is partitioned randomly in N folds so that Z will combine a group of folds i.e. $Z= F_1 \cup F_2 \cup F_3 \cup \ldots \cup F_N$. For each fold in Z, we train the data model over the OPF classifier and using the fitness function from one fold, which gets the best result to evaluate another fold to guide the optimization algorithm for selecting the best features.

A string of bits is associated with each particle of the population that shows the appearance of the feature as exist or not. The training fold F_i with its selected features let F_i^* is used to construct the classifier at this point only to assign the accuracy over other fold F_j as a fitness function. In the last part of the previous process, the whole population values were calculated and the particle with the highest value of fitness is denoted the best solution set with best features. As a result the training set with the selected features F^* is used to build the classification model and each solution through the classification process will be evaluated over the test set that is built over the specified folds in $Z \setminus \{ F_i \cup F_j \}$, this process is repeated for each fold in the dataset to be part of the training set and so we have N*(N -1) combinations which will be used for comparisons (Rodrigues et al., 2015).

Due to random partitioning, it may be classes with different sizes and the problem is the classifier always goes through the label with the highest class, therefore the accuracy will be down for the other classes with the lower values. To avoid this, the accuracy is measured by taking into consideration the classes with different sizes in the testing set F_j as shown in equation 15.

$$e_{i,1} = \frac{FP_i}{|F_j|-|F_j^i|} \qquad \textbf{(15)}$$

And

$$e_{i,2} = \frac{FN_i}{|F_j^i|} \text{i=1, 2… c} \qquad \textbf{(16)}$$

Where c is the number of classes, $|F_j^i|$ denotes the number of samples in fold F_j ,FP_i and FN_i is the false positives and false negatives respectively for class i, meaning while $FN_i$is the number of samples from class i and were classified as being from other classes in F_j(classified incorrectly), and FP_i is the number of samples from other classes but classified as being from the class i in F_j (classified correctly). The total error from class i will be defined using the error terms $e_{i,1}, e_{i,2}$ as shown in equation (17). Figure 2 describes briefly the evaluation technique.

$$E_i = e_{i,1} + e_{i,2} \qquad \textbf{(17)}$$

And at the end the accuracy Acc is calculated as follows:

$$Acc = 1 - \frac{\sum_{i=1}^{c} E_i}{2c} \qquad \textbf{(18)}$$

Construction of protein- protein interaction (PPI)

After reaching the best genes from our system, we want to search about protein–protein interactions to understand more about the genomic data and relation between these proteins. Based on the relationship between the genes and the matching probes, the probe- level of data GPL96 was used to simplify the process, as all probes have an expression value and we discovered that SLC5A1 gene is the protein which has the best features in our system results. Finally, to get the PPI the STRING database (Searching Tool for the Retrieval of Interacting Genes/Proteins) (http://string-db.org) is used to search about that gene and get the following protein- protein interaction as shown in Figure (3).

The figure shows the interaction between SLC5A1 and other genes and each one has a specific role to do in the process. Each node in the PPI network represents all the proteins produced by a single, protein-coding gene locus. The edges represent protein-protein associations, and fitness of edge in the PPI network indicates the strength of interaction between them. Associations between proteins in the network are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function, and this does not necessarily mean they are physically binding each other.

Table 1: Description of the datasets

Dataset	# Samples	# Tumor Samples	# Normal Samples	# Classes
TCGA	187	97	90	2
GSE10072	107	58	49	2

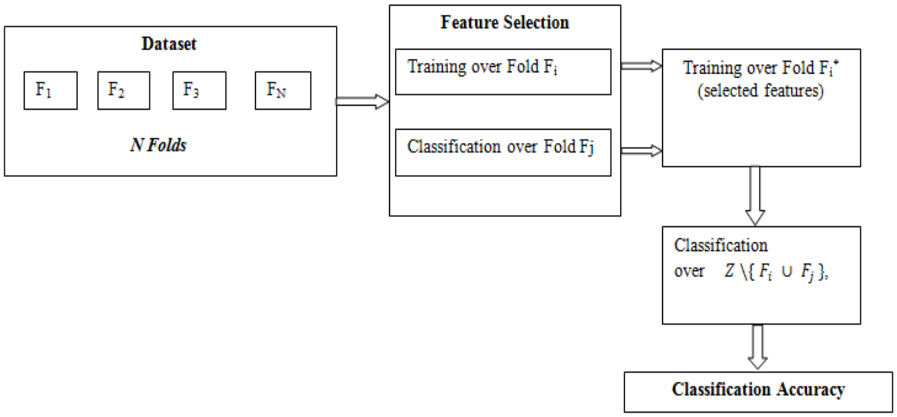
Dataset
F_1
F_2
F_3
F_N
N Folds
Feature Selection
Training over Fold F_i
Classification over Fold Fj
Training over Fold F_i^* (selected features)
Classification over $Z \setminus \{F_i \cup F_j\}$,
Classification Accuracy

Figure 1: Evaluation Technique for the Proposed Methodology

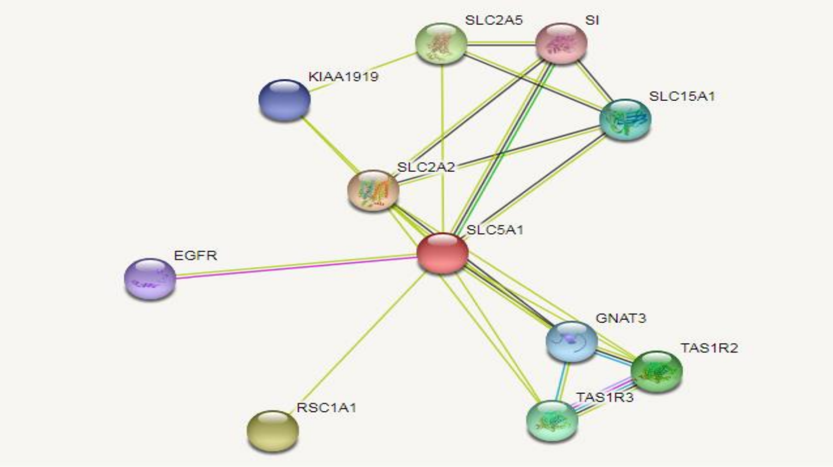

Figure 1: Protein-Protein interaction network for SLC5A1 from the STRING website, colored lines between the proteins indicate the various types of interaction evidence. Large protein nodes indicate the availability of 3D protein structure information.

The PPI showed the interaction between SLC5A1 with EGFR, which is defined as an important gene in the mutation process, as the mutation in this gene will lead to a production of a protein that is turned on or be activated; consequently, cells in the lung are signaled to constantly reproduce other cells leading to tumor formation and so lung cancer develops (Xu et al., 2016 and Powell et al., 2003). Besides other genes like GNAT3, Tas1R3, FOXK1, each one has a specific role to do in the community. The EGFR pathway is the main signaling pathway of lung cancer, and the mutation rate of its genes reaches 70-80% (Zhang et al., 2010).

RESULTS

The experimental results of the system were discussed regarding the first part of the proposed approach for the feature selection task. The optimization algorithms FPA, DEA, and hybrid system of FPA with DEA were implemented in R language (Nagarajan et al., 2013) on a Windows 7 operating system with 4GB RAM. Table 2 presents the parameter settings for the proposed hybrid approach.

Table 2: Parameter Settings used for each technique

Technique	**Parameters**
FPA	$\beta=1.5$,P= 0.8 , ʎ = 1.5, α =0.1
DEA	F =0.5, CR=0.8

Figure (4) displays Average convergence for the OPF accuracy over number of iterations for Dataset1 (TCGA) and dataset2 (GSE10072). It can be observed that by hybridization we can improve the results of the feature selection problem rather than executing each algorithm individually. All techniques used here can achieve a result of the feature selection, but the results showed that the hybrid technique is more suitable and efficient for feature selection tasks.

The statistical test of Wilcoxon Signed-Rank **(Wilcoxon)** was also performed to verify if there is a significant difference between the hybrid system and each technique used individually (Wilcoxon 1945). Table 3 displays the p-values of each system over the datasets and trained through over partitioning into folds. The bold values indicate whether there is a statistical difference with a significant level of α= 0.05 between the hybrid technique and the other techniques. It can be observed that there is a statistical difference between FPA and DEA for TCGA (Fold1) dataset. Similarly there is a difference between DEA and Hybrid approach for TCGA (Fold2) dataset and for TCGA (Fold3) between hybrid approach and FPA and DEA.

Furthermore, there is a statistical difference between FPA and DEA and Hybrid approach for GSE10072 (Fold1) and GSE10072 (Fold2) datasets.

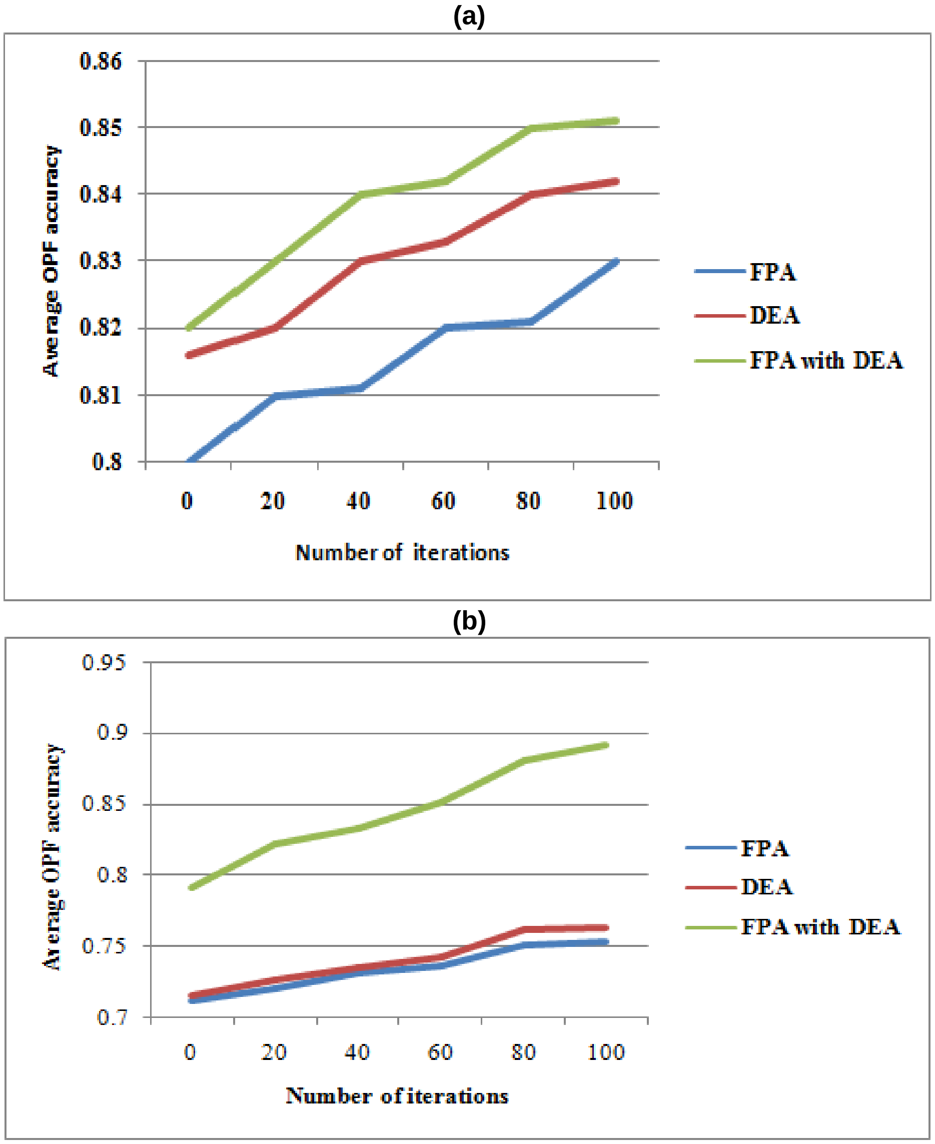

Figure 2: Average convergence for the OPF accuracy over number of iterations for Dataset1 (TCGA) and dataset2 (GSE10072)

Table 3:P-values of Wilcoxon Rank Test between FPA, DEA and the hybrid system, the bold values indicates if there is a statistical difference between the hybrid technique and the individual algorithms.

Dataset	FPA	DEA	FPA with DEA
TCGA (fold1)	0.8969819	0.2057047	0.0879398
TCGA (fold2)	0.01316537	0.05172055	0.0818493
TCGA (fold3)	0.05279535	0.2118896	0.0043550
GSE10072(Fold1)	0.3553007	0.76615579	0.9200293
GSE10072(Fold2)	0.01129972	0.3505584	0.8900574

Finally; it's clear that there is a statistical difference between the techniques which can be described as a little difference in the datasets.

DISCUSSION

This section discusses the experimental results of the proposed hybrid approach. Although some techniques already achieve good results, their deterministic characteristic might prevent the feasibility of the algorithm in complex situations. Thus, this paper purpose the use of (FPA, DEA, and Hybrid approach) as nondeterministic evolutionary algorithms to optimize the results. In order to get as accurate results as possible, the results presented in this paper are based on the average of 20 successive runs. The proposed algorithm seems to be efficient and robust. The presented algorithms (FPA, DEA, Hybrid approach) were implemented in R language. Moreover, each technique has been run 100 iterations over the attempted datasets.

To guarantee the selection of the best features in each iteration, the selected features of the solution that achieved the best accuracy are stored in a vector. Subsequently, the selected features in this vector will be used to authenticate the unknown features. Finally, the selected features that achieved the best precision will be used in the final step. Table 4 displays the accuracy results of each technique over the two datasets. The best values are formatted in bold for each technique and as it is shown in the table. It is clear that the hybrid approach achieved the best accuracy results in TCGA (Fold 1&2) and GSE10072 (Fold 1&2) datasets. Additionally, DEA had the best values in TCGA (Fold 3). It can be observed that although each algorithm could improve the results independently, the hybrid algorithm led to better results.

Table 4: Average accuracy for the different techniques over the datasets

Dataset	FPA	DEA	Hybrid Approach
TCGA (Fold1)	95.08	95.19	**95.32**
TCGA (Fold2)	95.13	95.18	**96.01**
TCGA (Fold3)	74.67	**76.50**	76.32
GSE10072(Fold1)	72.54	73.06	**74.55**
GSE10072(Fold2)	55.31	57.09	**58.93**

As a step towards the validation of the proposed approach, a PPI network is constructed using the STRING database for the processed genes resulting from the feature selection part. The results of PPI showed a clear interaction between SLC5A1 and EGFR. According to (Zhang et al., 2010), EGFR is defined as an important gene in the mutation process because its mutation leads to the production of a protein that signals the lung cells to reproduce constantly leading to tumor formation, and consequently lung cancer development. This study attempts to explore the effectiveness of using hybrid algorithms of feature selection and using it in cancer analysis, hence predicting related genes of this cancer type.

CONCLUSION

In this paper, a hybrid approach was optimized using FPA and DEA for solving feature selection problem of lung cancer and to discover key biomarkers and PPI for the gene expression data of the lung cancer. The experiments are implemented on two different datasets and the results showed that the hybrid approach possesses better results compared to working individually. Moreover, the SLC5A1 gene was discovered as a related gene of lung cancer. The PPI of SLC5A1 showed the interaction between EGFR and three other genes, namely GNAT3, Tas1R3, FOXK1 and others that have specific biological roles related to lung cancer. This study attempts to explore the effectiveness of using hybrid algorithms of feature selection and using it in cancer analysis, hence predicting related genes of this cancer type.

For future work, the hybrid approach will be applied to many other datasets in order to improve classification efficiency; it will be used with other classifiers like artificial neural network (ANN) and SVM. Furthermore, other Evolutionary algorithms can be used with either FPA or DEA to check the ability of hybridization to model the prediction process and to get the best results.

CONFLICT OF INTEREST

The authors declare no conflict of interest regarding this study.

AUTHOR CONTRIBUTIONS

All authors contributed in collecting and analyzing data. All authors participated in writing every part of this study. All authors read and approved the final version.

REFERENCES

Abdel-Raouf, O., Abdel-Baset, M., El-Henawy, I. (2014). An Improved Flower Pollination Algorithm with Chaos. *International Journal of Education and Management Engineering (IJEME), 4(2)*, 1-8.

Abraham A., Das S., Roy S. (2008). Swarm Intelligence Algorithms for Data Clustering. In: Maimon O., Rokach L. (eds), Soft Computing for Knowledge Discovery and Data Mining. Springer, Boston, MA, 279-313

Akutekwe, A., Seker, H., Iliya, S. (2014). An optimized hybrid dynamic Bayesian network approach using differential evolution algorithm for the diagnosis of Hepatocellular Carcinoma. *IEEE 6th International Conference on Adaptive Science & Technology (ICAST).*

Cai, Y., Du, J. (2014). Enhanced differential evolution with adaptive direction information. *IEEE Congr Evol Comput (CEC)*, Beijing, 305-312.

Coello, C., Lamont, G., Van Veldhuizen, D. (2015). Multi-objective Evolutionary Algorithms in Real-World Applications: Some Recent Results and Current Challenges. In Springer International Publishing, D.Greiner et al. (eds), *Advances in Evolutionary and Deterministic Methods for Design, Optimization and Control in Engineering and Sciences* (pp. 3-18).

Das, S., Suganthan, P.N. (2011). Differential evolution: A survey of the state-of-the-art. *IEEE Transactions on Evolutionary Computation 15*, 4-31.

Dorigo, M., Birattari, M., Stutzle, T. (2006). Ant colony optimization: Artificial ants as a computational intelligence technique. *IEEE Computational Intelligence Magazine*, 28-39.

Gong, W., Cai, Z., Ling, C. X. (2010). DE/ BBO: A hybrid differential evolution with biogeography based optimization for global numerical optimization. *Soft Computing 15*, 645-665.

Guyon, I., Weston, J., Barnhill, S. (2002). Gene selection for cancer classification using support vector machines. *Machine learning, Vol. 46(1-3)*, (PP 389-422).

He, X., Zhang, Q., Sun, N., Dong, Y. (2009). Feature selection with discrete binary differential evolution. In *Proceedings of the International Conference on Artificial Intelligence and Computational Intelligence* IEEE, Shanghai, (pp. 327-330).

John, G.H., Kohavi, R., Pfleger, K. (1994). Irrelevant features and the subset selection problem. *ICML 94*, (pp. 121-129).

Kancherla K., Mukkamala S. (2012) Feature Selection for Lung Cancer Detection Using SVM Based Recursive Feature Elimination Method. In: Giacobini M., Vanneschi L., Bush W.S. (eds) Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics. EvoBIO 2012. Lecture Notes in Computer Science, vol 7246. Springer, Berlin, Heidelberg (pp.168-176).

Karaboga, D., Basturk, B.J (2007). A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *Journal of Glob Optim vol. 39, no. 3*, 459-471.

Nagarajan, R., Scutari, M., Le'bre, S. (2013). *Bayesian Networks in R with Applications in Systems Biology.* Springer-Verlag New York, (pp. XIII, 157).

Nakamura, R. Y., Pereira, L. A., Costa, K. A., Rodrigues, D., Papa, J.P., Yang, X.S., (2012).BBA: A binary bat algorithm for feature selection. In *Proceedings of 25th the Conference on Graphics, Patterns and Images (SIBGRAPI)*, IEEE, Ouro Preto, Brazil, (pp. 291-297).

Papa, J. P., Falcao, A. X., Suzuki, C. T. (2009). Supervised pattern classification based on optimum-path forest. *Int. J. Imaging Syst. Technology. 19 (2)*, 120-131.

Papa, J. P., Falcao, A. X., Albuquerque, V. H., Tavares., J. M. (2012). Efficient supervised optimum-path forest classification for large data sets. *Pattern Recognition*, 512-520.

Powell, C., Spira, A., Derti, A. (2003). Gene expression in lung adenocarcinoma of smokers and nonsmokers. *AJRCMB vol 29*, 157-162.

Qin, A.K., Suganthan, P.N. (2005). Self-adaptive differential evolution algorithm for numerical optimization. *Proc. IEEE Congress on Evolutionary Computation, Edinburgh, Scotland, pp. 1785-1791.*

Qin, A.K., Huang, V.L., Suganthan, P.N. (2009). Differential Evolution Algorithm with Strategy Adaptation for Global Numerical Optimization. In *IEEE Transactions on Evolutionary Computation*, vol. 13, no. 2, pp. 398-417

Retrieved Jan 206, from the cancer genomic Atlas website: https://cancergenome.nih.gov/

Robic, T., Filipic, B. (2005). DEMO: Differential evolution for multi objective optimization. In *Evolutionary Multi-Criterion Optimization.* Springer Berlin Heidelberg. (pp. 520-533).

Rodrigues, D., Yang, X. S., Souza, A., Papa, J. (2015). Binary flower pollination algorithm and its application to feature selection. In *Recent Advances in Swarm Intelligence and Evolutionary Computation Studies in Computational Intelligence* .Springer International Publishing (pp. 85-100).

Sayed, S.A., Nabil, E., Badr, A., (2016). A binary clonal flower pollination algorithm for feature selection, *Pattern Recognition Letters*: North-Holland (pp. 21-27).

Sikdar, U. K., Ekbal, A., Saha, S. (2012). Differential evolution based feature selection and classifier ensemble for named entity recognition. *In COOLING*, 2475-2490.

Storn, R., Price, K. (1997). Differential evolution-A simple and efficient heuristic for global optimization over continuous Spaces. *J. Global Optim., vol. 11* , 341-359.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., (2011). The STRING database functional interaction networks of proteins globally integrated and scored. *Nucleic Acids Res 39 (561- 568).*

Wilcoxon, F. (1945). Individual comparisons by ranking methods. Biometrics Bulletin 6(1), 80-83.

Xu, H., Ma, J., Wu, J., Chen, L., Sun, F., Qu, C., Xu, S. (2016). Gene expression profiling analysis of lung adenocarcinoma. Brazilian journal of medical and biological research 49(3), e4861. Doi: 10.1590/1414-431X20154861

Yang, X. S. (2012). Flower pollination algorithm for global optimization. In *International conference on unconventional computing and natural computation*, Springer, Berlin Heidelberg (pp. 240-249).

Yang, X.S., Karamanoglu, M., He, X. S. (2013). Multi-objective Flower Algorithm for Optimization. *Procedia Computer Science, vol. 18*, 861-868.

Zhang, W. J., Xie, X. F. (2003). DEPSO: hybrid particle swarm with differential evolution operator. *IEEE Int. Conf. on Systems, Man and Cybernetics, Washington DC, USA, Vol. 4*, 3816-3821.

Zhang, X., Zhang, S., Yang, X., Yang, J., Zhou, Q., Yin, L., An, S., Lin, J., Chen, S., Xie, Z., Zhu, M., Wu, YL. (2010). Fusion of EML4 and ALK is associated with development of lung adenocarcinomas lacking EGFR and KRAS mutations and is correlated with ALK expression. *Molecular Cancer 9:188*, Doi: 10.1186/1476-459.