Feature Selection using Salp Swarm Algorithm for Real Biomedical Datasets

Hadeel Tariq Ibrahim[†,††] , Wamidh Jalil Mazher[†,†††] , Osman N. Ucan1[†] ,Oguz Bayat[†]
[†] Altinbas University, Department of Electrical and Electronics Engineering, Istanbul, Turkey
[††] Diyala University, College of Basic Education, Diyala, Iraq
[†††] Southern Technical University, Basrah, Iraq

Summary
The main objective of this paper is to develop a new powerful heuristic optimization algorithm to be used in feature selection. Here, the use of Salp Swarm Algorithm in feature selection (SSA-FS) is proposed for the first time in literature. SSA-FS has been compared with Particle Swarm Optimization and Differential Evolution performance with criteria of accuracy and runtime. In this paper, real datasets obtained from Iraqi hospitals for breast, bladder and colon cancers and synthetic datasets for evaluation. We have found that SSA-FS has been achieved the highest accuracies with less runtime in comparison with other selected algorithms for both real and synthetic datasets.

Key words:
Feature selection, Salp Swarm Algorithm, Particle Swarm Optimization, Differential Evolution

1. Introduction

Applying for feature selection improves classification performance by deleting irrelevant and redundant datasets attributes. It reduces training time and confronts the curse of dimensionality[1]. There are many heuristic optimization algorithms have been employed in feature selection, such as Genetic Algorithm (GA) [3], Differential Evolution(DE) [4]–[7], Particle Swarm Optimizer (PSO)[8], Ant Colony Optimization(ACO)[9], Grey Wolf Optimization (GWO)[10] and Moth Flame Optimization[11], Multiverse Optimizer(MVO)[12].

We developed Salp Swarm Algorithm (SSA) to establish the proposed approach, SSA-FS. Authors approved the high performance of SSA-FS by comparing it with another algorithms. The other algorithms like Particle Swarm Optimization (PSO)[8] and Differential Evolution (DE)[4]. Salps are creatures like jellyfishes but living as swarms in deep oceans. Salps swarms are moving by water forces to find food, they organized as salp chains with head salp and followers[13].

The proposed approach, SSA-FS, has been examined on real biomedical datasets for breast, bladder and colon cancers in Iraq for (2010-2012) period as mentioned in Table 1. The lowest runtimes have been obtained from the SSA-FS approach with employing all datasets. To assess SSA-FS performance and approve its efficiency, we compared it with another two algorithms, Particle Swarm Optimizer (PSO-FS) and Differential Evolution (DE-FS). The novelty of the current paper is apparent in developing SSA to be applied in feature selection, in our knowledge; this is the first time for employing SSA in feature selection. This paper is organized as follows: next section explains Particle Swarm Optimizer (PSO) and Differential Evolution (DE). Salp Swarm Algorithm (SSA) composition demonstrated in section 3. Section 4 explained proposed SSA-FS. Results discussing and analyzing has shown in section 5. Finally, the conclusion and future works have presented in section 6.

2. Particle Swarm Optimization (PSO) and Differential Evolution (DE)

Particle Swarm Optimization (PSO) is a population-based algorithm inspired from birds lives, specifically their movements in swarms to find food[14]. It based on using a number of particles that compose a swarm which is moving around in the search space and searching for the best solution. PSO has used the following variables:

P_a: Population of agents
ag_i: ith agent
pl_i: ag_i location in solution space
O_f: Objective function
vc_i: ag_iagent velocity
$VC(ag_i)$: ag_iagent neighbourhood (specific)

Let us explain the PSO mathematical model which based firstly on particle amending rule:

$$l = l + rd \quad (1)$$

With:

$$rd = rd + c_l * rnd * (l_{Best} - l) + c_g * rnd * (s_{Best} - l) \quad (2)$$

Where:
l: particle's location
rd: route direction
c_l: local information weight
c_g: global information weight
l_{Best}: particle's best location
s_{Best}: best location of the swarm
rnd: random parameter

Manuscript received December 5, 2017
Manuscript revised December 20, 2017

c_l and c_g are important to specify personal best value and neighborhood best value respectively. There are three powers affect in PSO results, inertia, personal power and social power. Inertia obliges the particle to move in similar direction with equal velocity. Personal power encourages the particle to turn back is the previous location is better than existing one. Lastly, social power forces the particle to keep track of the best adjacent direction, eq. (3).

$$vc_i^{t+1} = \underbrace{vc_i^t}_{Inertia} + \underbrace{c_l U_1^t (l_{Best(i)}^t - l_i^t)}_{Personal\ power} + \underbrace{c_g U_2^t (s_{Best}^t - l_i^t)}_{Social\ Power} \quad (3)$$

In spite of PSO quick convergence, but it suffers from dropping into local optima especially in high search space.

Differential Evolution (DE)[15] it is an evolutionary, population-based algorithm. Any evolutionary algorithm passed in four steps, initialization, mutation, recombination and selection. Determining the parameters' upper and lower bounds applied in initialization: $a_j^L \le a_{j,i,1} \le a_j^U$ where initial values must be in interval [a_j^L, a_j^U]. In mutation, arbitrarily three vectors have chosen, a_{r_1}, G_p, a_{r_2}, G_p and a_{r_3}, G_p. In Eq. (4) the difference of two weighted vectors is added to the third.

$$v_i, G_p + 1 = a_{r_1}, G_p + M_F(a_{r_2}, G_p - a_{r_3}, G_p) \quad (4)$$

Where M_F is the mutation factor and $v_i, G_p + 1$ is the donor vector. The trial vector $u_i, G_p + 1$ is created in recombination step as shown in Eq. (5).

$$u_{j,i}, G_p + 1 = \begin{cases} v_{j,i}, G_p + 1 & if\ rnd_{j,i} \le P_r\ or\ j = I_{rnd} \\ a_{j,i}, G_p & if\ \ rnd_{j,i} > P_r\ or\ j \ne I_{rnd} \end{cases} \quad (5)$$

Where P_r is the probability and rnd is random numbers.

Finally, in selection step, the lowest value is considered between the target vector a_i, G_p and the trial vector $v_i, G_p + 1$, as shown in Eq. (6).

$$a_i, G_p + 1 = \begin{cases} u_i, G_p + 1 & if\ \ f(u_i, G_p + 1) \le f(a_i, G_p) \\ a_i, G_p & otherwise \end{cases} \quad (6)$$

The previous steps are repeated until satisfied the stopping condition. DE is easy to use and choose their parameters but cannot ensure the global solution because it can stack at local optima especially in high dimensional space.

3. Salp Swarm Algorithm (SSA)

Salps are part of Salpidae family with the limpid cylinder-design body, they look like jellyfishes in texture and movement. The water is pushed salps bodies to progress[16]. Salps swarming attitude is the main inspiration to build Salp Swarm Algorithm[13]. Salps compose a swarm in profound oceans; this swarm named salp chain. The cause of salps swarm behavior is not well expressed yet, nevertheless some researchers consider such behavior has been done to enhance their movement in seeking for food[17].

Originally, the salps population has been divided into two groups: head and followers to formulate the mathematical model for salp chains. The head position is at the beginning of the chain, where the remainder is the followers. Like the other swarm inspired optimization approaches, N is the number of problems variables where the salp location is determined in N-dimensional space of searching. Accordingly, the salps locations are saved in a matrix with two dimensions named Xl. The food place is the goal of salp swarms, called FP[13].

Equation 2.1 is suggested to modify the head location:

$$Xl_i^1 = \begin{cases} FP_i + r_1((U_i - L_i)r_2 + L_i) & r_3 \ge 0 \\ FP_i - r_1((U_i - L_i)r_2 + L_i) & r_3 < 0 \end{cases} \quad (7.1)$$

Where Xl_i^1 shows the head location in i_{th} dimension, the place of food in ith dimension is represented by FP_i . The upper and lower bounds are shown as U_i and L_i respectively, r_1, r_2 and r_3 are random numbers. Only the head of salp chain has the right to modify its location relative to food place, this fact is clear in Eq. (7.1). The most effective parameter in SSA is r_1 which makes the exploration and exploitation phases in balanced state, this is shown in Eq. (7.2):

$$r_1 = 2e^{-\left(\frac{4t}{T}\right)^2} \quad (7.2)$$

Where t is the present iteration and T is the total iterations. The variables r_2 and r_3 are arbitrarily created in the period [0,1], they determine if the next location in ith dimension must be in positive or negative infinity in addition to the pace size. Eq. (7.3) is used to modify the follower's locations (Newton's law of motion):

$$Xl_i^j = \frac{1}{2}\sigma tm^2 + vc_0 tm \quad (7.3)$$

Where j≥2 and Xl_i^j is the location of jth followers of salps in ith dimension, tm is time, vc_0 is the initial velocity, and $\sigma = \frac{vc_{final}}{vc_0}$ where $vc = \frac{Xl - Xl_0}{tm}$. Since the optimization time is iteration, the difference between iterations is 1, Eq. (7.3) can be formulated as following where $vc_0 = 0$:

$$Xl_i^j = \frac{1}{2}(Xl_i^j + Xl_i^{j-1}) \quad (7.4)$$

Xl_i^j presents the jth follower location of salp chain where j≥2 in ith dimension.

4. Proposed SSA-FS

The required issues to build SSA-FS paradigm are listed below:

4.1 Encrypted plan

We tend to encrypt the individuals by employing a vector of real numbers. The vector is used for features which mapped arbitrarily to be in [0,1] interval as shown in Fig .1 upper part. Accordingly, if the component value is equal to or more than 0.5, it will be substituted with 1 so the feature is chosen, otherwise, the value approximated to 0 and the feature is not chosen, as shown in Fig .1 lower part:

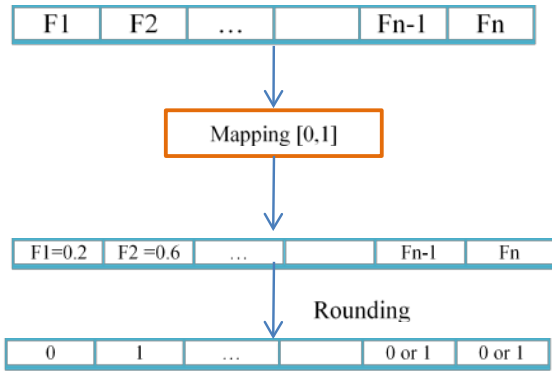

Figure 1. Data mapping and decoding

4.2 Objective function

Our objective function based on calculating accuracy for each selection, accuracy is calculated by Eq. (8):

$$Acc = \frac{T_P + T_N}{T_P + F_N + F_P + T_N} \quad (8)$$

Where:

T_P: is the number of correct predictions and actual class is true.

T_N: is the number of correct predictions and actual class is false.

F_N: is the number of incorrect predictions and actual class is true.

F_P: is the number of incorrect predictions and actual class is false.

4.3 System architecture

We described our proposed system, SSA-FS architecture in this section. Previous studies used the term 'System Architecture' [18], [19]. The main parts of SSA-FS are:

- Data normalization: is a common preprocess in selecting features. We normalized the features to exist in [0,1] interval and avoid the bad effect of existing some bias values of some features, this normalization has applied by determining the selected feature by FB in Eq. (9):

$$FB = \frac{FA - min_{FA}}{max_{FA} - min_{FA}} \quad (9)$$

- Salps individuals decoding: in this step, our vector has been occupied by the selected features.
- Determining training and testing sets: now we divided the dataset into, training set (X_{train},Y_{train}) and testing set (X_{test}, Y_{test}). As shown in left part of Fig .2, the main features have been represented by X1, X2,… and the main class is Y. To generate the model, we managed X_{train} and Y_{train} by applying any classifier like SVM. We entered X_{test} as input for the model to examine its accuracy and Y? output. Ground truth is obtained if Y? equals Y_{test} as shown in the right part of Fig. 2. Finally, we test the model accuracy by employing X_{test} as input to the model and the output has obtained from the model named Y? to compare it with Y_{test} , if they are equal, this output is the ground truth.

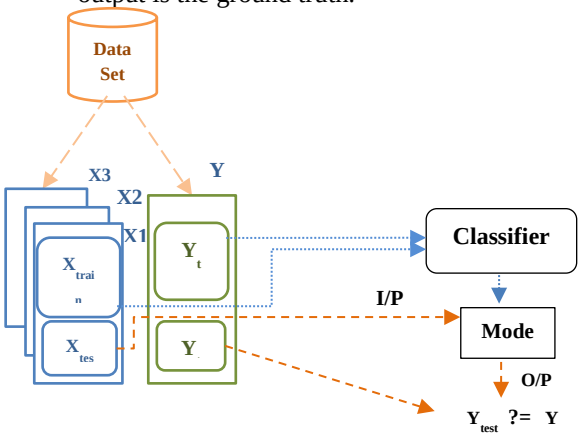

Figure 2. Determining training and testing sets process

- Choose features subset: from the training set, we chose features with 1's value.
- Fitness assessment: to learn our classifier, we utilized training set vectors and then determined classification accuracy using Eq. 8.
- Termination condition: we stopped the whole process by setting the maximum iteration.

The entire system workflow for SSA-FS is clarified in Fig. 3, which shows the relationships among main system parts.

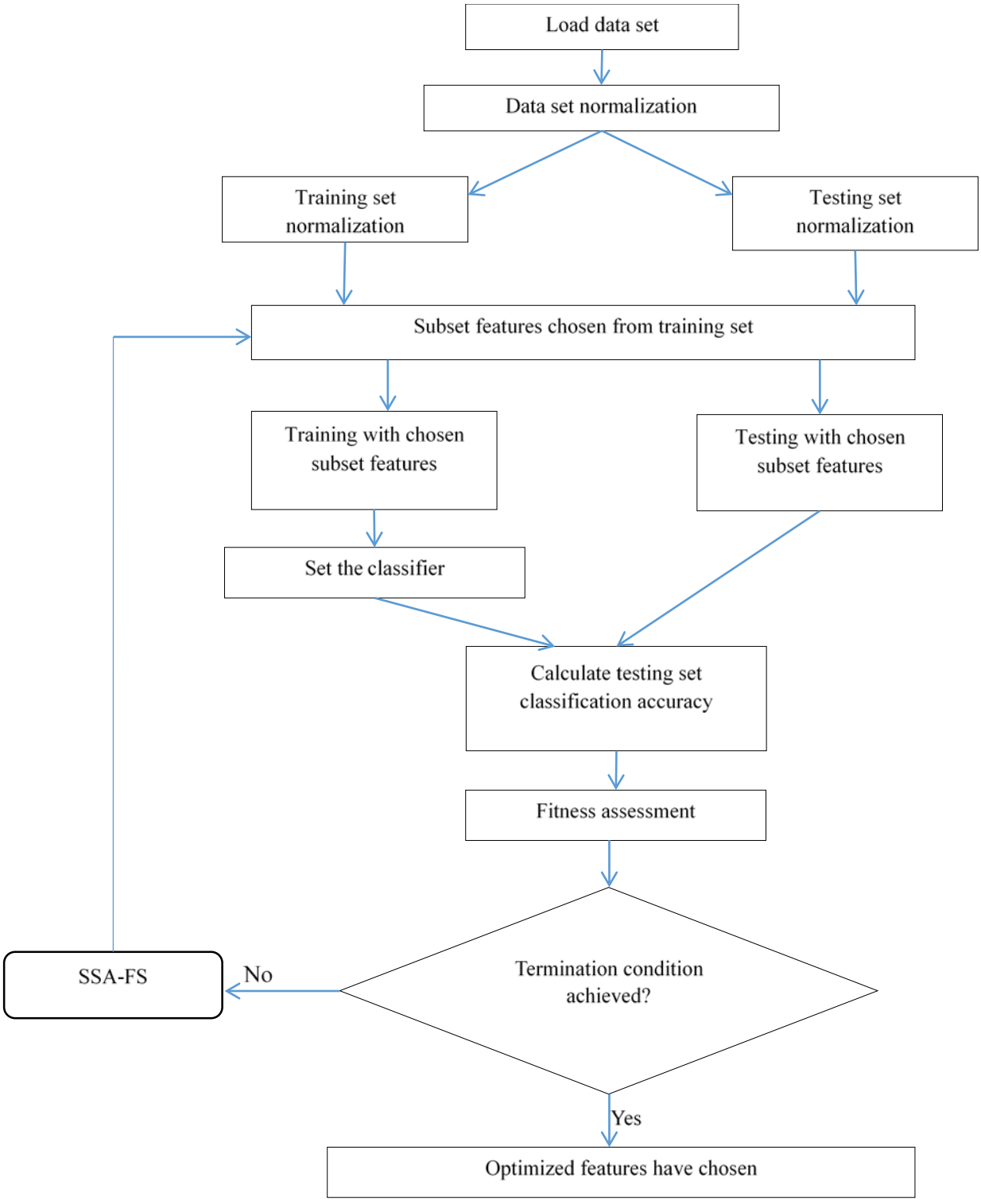

Figure 3. SSA-FS flowchart

5. Results

SSA-FS applied on a personal portable computer with Intel(R) Core(TM) i7-5500U CPU, 2.40 GHz, 8 GB RAM and Windows 10 as the operating system. We utilized Matlab R2015a to implement our research.
In the current research, we examined two kinds of datasets, real and synthetic biomedical datasets, such datasets are listed in Table 1 and Table 2.

Table 1. Real biomedical datasets[20]

Datasets	Dataset's year	No. of instances	No. of features
Breast Cancer	2010	3151	16
	2011	3683	16
	2012	3836	16
Bladder Cancer	2010	1301	16
	2011	1530	16
	2012	1457	16
Colon Cancer	2010	906	16
	2011	1135	16
	2012	1217	16

Table 2. Synthetic biomedical datasets

Datasets	No. of instances	No. of features
Breast Cancer[21]	683	11
Bladder Cancer(Biostat 514/517 Datasets, n.d.)	2922	9
Colon Cancer[22]	1858	16

We employed official biomedical real datasets for breast, bladder and colon cancers in Iraqi hospitals for (2010-2012) period. There were some noises in these datasets, so they have been cleaned up and removed bias and irregular values that effect on classification performance. The SSA-FS approach has been achieved with high performance employing all real datasets. Authors compared SSA-FS results with results of two more approaches like DE-FS and PSO-FS utilizing breast, bladder and colon cancers in Iraq datasets. Such comparisons are based on two criteria, the runtime, and accuracy. Table 3 lists the results for SSA-FS compared with DE-FS and PSOFS algorithms applied on real datasets for breast, bladder and colon cancers in Iraq for (2010-2012) period.
The results listed in Table 3 are visualized in figures 4 and 5. Figure 4 shows results considering accuracy criteria where figure 5 browses result considering runtime criteria.

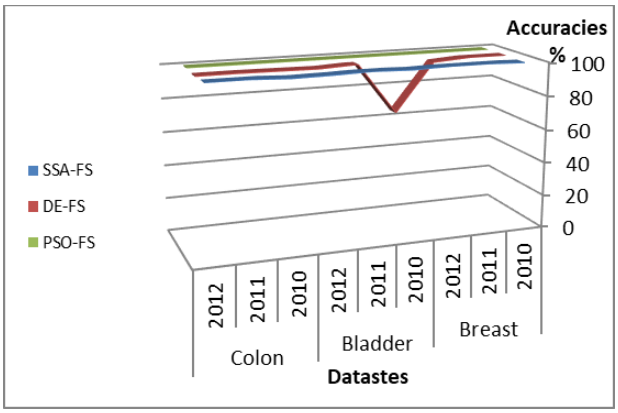

Figure 4. Results for SSA-FS, DE-FS, and PSO-FS considering accuracies for real datasets (breast, bladders and colon between 2010-2012 period in Iraq)

As shown in figure 4, SSA-FS and PSO-FS have been achieved highest and closest accuracies (99%-100%) utilizing the real datasets. In the other hand, DE-FS has been obtained accuracies between (70%-100%).

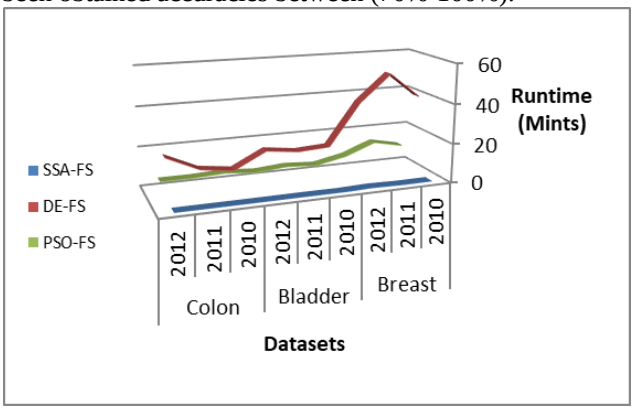

Figure 5. Results for SSA-FS, DE-FS, and PSO-FS considering runtimes for real datasets (breast, bladders and colon between 2010-2012 period in Iraq)

For the same datasets and by conducting the previous three approaches, the authors calculated the runtimes, as visualized in figure 5. Here, we approved the high convergence rate of SSA to optimize the selection of features, SSA-FS runtimes were less than one minute (several seconds) for all specified datasets. PSO-FS comes in the second order after SSA-FS where it spent several minutes to be executed; finally, DE-FS needs more runtime near to one hour sometimes.

To evaluate our proposed approach, SSA-FS, we re-applied SSA-FS, PSO-FS, and DE-FS on synthetic datasets for breast, bladder and colon cancers. Results are listed in table 4.

Table 3. Results for SSA-FS, DE-FS and PSO-FS algorithms utilizing real datasets for breast, bladder and colon cancers in Iraq for (2010-2012) period

Dataset name	Datasets years	Algorithms	Best Accuracy %	No. of selected features	Run Time (Mints)
Breast Cancer	**2010**	**SSA-FS**	99.33	6	0.30
		DEFS	100	7	40.26
		PSO-FS	99.94	5	10.17
	2011	**SSA-FS**	100	6	0.22
		DE-FS	100	7	52.8
		PSO-FS	99.92	6	13.36
	2012	**SSA-FS**	100	6	0.55
		DEFS	98.44	7	39.52
		PSO-FS	99.93	5	7.47
Bladder Cancer	**2010**	**SSA-FS**	99.33	6	0.09
		DE-FS	70.09	8	18.47
		PSO-FS	99.91	6	4.29
	2011	**SSA-FS**	100	6	0.16
		DE-FS	100	7	17.56
		PSO-FS	99.89	7	5.24
	2012	**SSA-FS**	99.33	6	0.11
		DE-FS	98.83	8	19.53
		PSO-FS	99.88	7	4.0
Colon Cancer	**2010**	**SSA-FS**	99	6	0.08
		DE-FS	99.05	5	11.53
		PSO-FS	99.88	6	5.30
	2011	**SSA-FS**	100	5	0.08
		DE-FS	99.25	7	13.59
		PSO-FS	99.86	6	4.56
	2012	**SSA-FS**	100	7	0.08
		DE-FS	99.14	5	21.01
		PSO-FS	99.87	6	4.55

Table 4. Results for SSA-FS, DE-FS and PSO-FS algorithms utilizing synthetic datasets for breast, bladder and colon cancers

Datasets	Algorithms	Best Accuracy %	No. of selected features	Runtime (Mints)
Breast Cancer	**SSA-FS**	98.75	5	0.03
	DE-FS	100	7	6.41
	PSO-FS	99.67	8	3.11
Bladder Cancer	**SSA-FS**	100	4	0.36
	DE-FS	77.01	6	68.40
	PSO-FS	99.75	6	4.26
Colon cancer	**SSA-FS**	99.75	5	0.04
	DE-FS	66.66	8	38.32
	PSO-FS	99.74	7	3.08

Figures 6 and 7 are visualizing table 4 contents. Figure 6 shows the re-applying of the specified approaches on synthetic datasets with considering accuracies, where figure 7 considers runtimes.

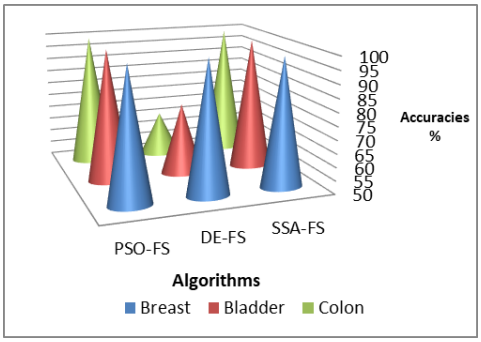

Figure 6. Results for SSA-FS, DE-FS, and PSO-FS considering accuracies for Synthetic datasets (breast, bladders and colon cancers)

Similar to the previous results (on real datasets), SSA-FS and PSO-FS have been gained the highest accuracies vs DE-FS which obtained the lowest accuracies.

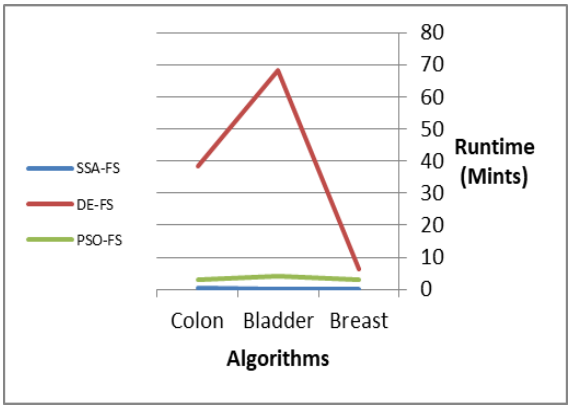

Figure 7. Results for SSA-FS, DE-FS, and PSO-FS considering runtimes for Synthetic datasets (breast, bladders and colon cancers)

Finally and again, SSA-FS has been gained the lowest runtime (near to zero). PSO-FS occupied the second order and the last one is DE-FS which spent more than one hour sometimes.

6. Conclusion

This paper proposed a new approach in feature selection by developing a recent heuristic optimization algorithm, Salp Swarm Algorithm (SSA). The new proposed approach named SSA-FS. Results from such approach compared with previous approaches like DE-FS and PSO-FS. The comparisons criteria are accuracy and runtime. We applied SSA-FS, DE-FS, and PSO-FS on two types of datasets, real and synthetic. We found that SSA-FS achieved the highest accuracies and lowest runtimes for all datasets vise DE-FS and PSO-FS. The main drawback of PSO- and DE approaches is the low convergence rate because of their ability to stack in local optima. In the other hand, we approved the high rate of convergence for SSA in selecting features in all datasets, such approval is clear in low runtimes of SSA-FS. We proposed for future work to employ SSA in another optimization approach in addition to the current paper approach, i.e. feature selection.

References

[1] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” J. Mach. Learn. Res., vol. 3, no. 3, pp. 1157–1182, 2003.

[2] R. Leardi, Amparo lupianez gonzales, and A. Lupiáñez González, “Genetic algorithms applied to feature selection in PLS regression: how and when to use them,” Chemometrics and Intelligent Laboratory Systems, vol. 41, no. 2. pp. 195–207, 1998.

[3] C. L. Huang and C. J. Wang, “A GA-based feature selection and parameters optimization for support vector machines,” Expert Syst. Appl., vol. 31, no. 2, pp. 231–240, 2006.

[4] R. N. Khushaba, A. Al-ani, A. Al-jumaily, and P. O. Box, “Differential Evolution based Feature Subset Selection,” Evolution (N. Y)., 2008.

[5] R. N. Khushaba, A. Al-Ani, and A. Al-Jumaily, “Feature subset selection using differential evolution and a statistical repair mechanism,” Expert Syst. Appl., vol. 38, no. 9, pp. 11515–11526, 2011.

[6] A. Al-Ani, A. Alsukker, and R. N. Khushaba, “Feature subset selection using differential evolution and a wheel based search strategy,” Swarm Evol. Comput., vol. 9, pp. 15–26, 2013.

[7] O. Ceylan and T. Gulsen, “A Comparison of differential evolution and harmony search methods for SVM model selection in hyperspectral image classification,” Igarss 2016, pp. 485–488, 2016.

[8] C. S. Yang, L. Y. Chuang, J. C. Li, and C. H. Yang, “Chaotic maps in binary particle swarm optimization for feature selection,” pp. 107–112, 2008.

[9] H. R. Kanan, K. Faez, and S. M. Taheri, “Feature Selection Using Ant Colony Optimization (ACO): A New Method and Comparative Study in the,” pp. 63–64, 2007.

[10] E. Emary, Hossam m.Zawbaa, C. Grosan, and A. E. Hassenian, Feature Subset Selection Approach by Gray-Wolf Optimization, vol. 334. 2015.

[11] H. M. Zawbaa, E. Emary, B. Parv, and M. Sharawi, “Feature selection approach based on moth-flame optimization algorithm,” Proc. IEEE Congr. Evol. Comput., pp. 4612–4617, 2016.

[12] A. M. et al. Faris, H., Hassonah, M.A., Al-Zoubi, “A multi-verse optimizer approach for feature selection and optimizing SVM parameters based on a robust system architecture,” Neural Comput Applic, vol. doi:10.100, 2017.

[13] S. Mirjalili, A. H. Gandomi, S. Z. Mirjalili, S. Saremi, H. Faris, and S. M. Mirjalili, “Salp Swarm Algorithm: A bio-inspired optimizer for engineering design problems,” Adv. Eng. Softw., no. July, 2017.

[14] J. Kennedy and R. Eberhart, “Particle swarm optimization,” Neural Networks, 1995. Proceedings., IEEE Int. Conf., vol. 4, pp. 1942–1948 vol.4, 1995.

[15] R. Storn and K. Price, “Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces,” J. Glob. Optim., vol. 11, no. 4, pp. 341–359, 1997.

[16] L. P. Madin, “Aspects of jet propulsion in salps,” Can. J. Zool., vol. 68, no. 4, pp. 765–777, 1990.

[17] P. A. V Andersont and Q. Bone, “Communication between individuals in salp chains II. Physiology,” Biol. Sci. Proc. R. Soc. Lond. B, vol. 210, no. 210, pp. 559–574, 1980.

[18] C. L. Huang and C. J. Wang, “A GA-based feature selection and parameters optimization for support vector machines,” Expert Syst. Appl., vol. 31, no. 2, pp. 231–240, 2006.

[19] H. Faris, M. A. Hassonah, A. M. Al-Zoubi, S. Mirjalili, and I. Aljarah, “A multi-verse optimizer approach for feature selection and optimizing SVM parameters based on a robust system architecture,” Neural Comput. Appl., pp. 1–15, 2017.

[20] Ministry of Health-Iraq-Iraqi Cancer Board, Acceptance of Official Cancer datasets from Iraq. 2017.
[21] Lichman M, "UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences.," 2013. [Online]. Available: http://archive.ics.uci.edu/ml. [Accessed: 01-Jul-2017].
[22] "Biostat 514/517 Datasets." http://courses.washington.edu/b517/Datasets/datasets.html.