iJournals: International Journal of Software & Hardware Research in Engineering
ISSN:2347-4890, ijournals.in/ijshre
Volume 12 Issue 4, April 2024

Feature Selection for Cancer Datasets by Modifying Marine Predator Algorithm

Author: Hadeel Tariq Ibrahim[1]; Wamidh Jalil Mazher[2]; Zainab Fadhil Yaseen [3]

Affiliation: Head of Information Technology Division, Al-Shatrah Univ. , Al-Shatrah, Iraq[1]; Electrical Engineering Dept., Southern Technical University, Basra, Iraq[2]; University of Thi-Qar, Nassirya, Iraq[3]

E-mail: Hadeel.tariq@shu.edu.iq[1]; wamidh.mazher@stu.edu.iq[2]; zyaseen@utq.edu.iq [3]

DOI:10.26821/IJSHRE.12.4.2024.120403

ABSTRACT

This paper's main goal is to modify an efficient heuristic optimization approach for feature selection purpose. Here, Marine Predator Algorithm for Feature Selection (MPAFS) is proposed. With regard to runtime and accuracy, MPAFS has been compared to Particle Swarm Optimization and Differential Evolution. Real datasets for breast, bladder, and colon cancers were gathered from Iraqi hospitals for this study, along with artificial datasets for assessment. For both actual and artificial datasets, we discovered that MPAFS attained the best accuracies with the shortest runtime when compared to other chosen methods.

Keywords: ***Feature selection (FS) , Marine Predator Algorithm (MPA), Particle Swarm Optimization Algorithm (PSO), Differential Evolution Algorithm (DE)***

1. INTRODUCTION

By removing pointless and superfluous dataset attributes, applying for feature selection enhances classification performance. It cuts down on training time and combats the dimensionality curse. [1]. Numerous heuristic optimization techniques, including the Genetic Algorithm, have been used in feature selection (GA)[2], Particle Swarm Optimizer (PSO)[3], Differential Evolution(DE) [4]–[7], Ant Colony Optimization(ACO)[8], Moth Flame Optimization[9], Multiverse Optimizer(MVO)[10], Harris hawk algorithm [11] ,Grasshopper Optimizer[12] and Grey Wolf Optimization (GWO)[13]

To produce the suggested method, MPAFS, we modified the Marine Predator Algorithm (MPA). By comparing MPAFS with extra algorithms (Particle Swarm Optimization (PSO)[3] and Differential Evolution (DE)[4], the authors were able to demonstrate the technique's excellent performance. The MPA algorithm is a revolutionary approach that utilizes the optimum encounter rate strategy in biological interactions between prey and predator as well as the Lévy and Brownian foraging tactics in ocean predators. [14].

The proposed method, MPAFS, has been tested using real cancer datasets for colon, bladder, and breast cancers in Iraq sk that authors follow some simple guidelines. In essence, we ask you to make your paper look exactly like this document. The easiest way to do this is simply to download the template, and replace the content with your own material.

2. PARTICLE SWARM OPTIMIZATION (PSO) AND DIFFERENTIAL EVOLUTION (DE)

A population-based technique called Particle Swarm Optimization (PSO) was inspired by how birds move in swarms to obtain food[13]. It is based on the use of a large number of particles that make up a swarm that moves across the search space looking for the best answer. The following variables were used by PSO:

P_a: Agents' Population

ag_i: i[th] agent

pl_i: ag_i location in solution space

O_f: Objective function

vc_i: ag_i agent velocity

$VC(ag_i)$: ag_i agent neighbourhood (specific)

The PSO mathematical model, which is based initially on the particle amendment rule, is explained as follows: $l = l + rd$ (1)

With:

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

Where:

l: particle's place

rd: route way

c_l: local info weight

c_g: global info weight

l_{Best}: particle's best place

s_{Best}: top place of the swarm

rnd: arbitrary parameter

c_l and c_g are important to specify personal best value and neighborhood best value respectively. It is crucial to provide the personal best deal and neighborhood best value using c_l and c_g, respectively. Inertia, personal power, and societal power are the three forces that have an impact on PSO outcomes. The particle must travel in the same direction and at the same speed due to inertia. When the old place is preferable to the current one, personal power promotes the particle to return there. The particle must also keep track of the ideal nearby direction, due to social power eq. (3).

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

PSO suffers from sliding into local optima despite its rapid convergence, particularly in large search space. The population-based evolutionary algorithm Differential Evolution (DE) uses evolution. The four steps of every evolutionary algorithm are initialization, mutation, recombination, and picking. Identifying the upper and lower boundaries for the parameters used for initialization [15]:

$a_j^L \le a_{j,i,1} \le a_j^U$ where initial values must be in interval $[a_j^L, a_j^U]$. Three vectors have been randomly selected in mutation, a_{r_1}, G_p, a_{r_2}, G_p and a_{r_3}, G_p. In Eq. (4) the third weighted vector is increased by the difference of the first two.

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

Where M_F is the mutation factor and $v_i, G_p + 1$ is the donor vector. The test vector $u_i, G_p + 1$ is bent in recombination step as revealed 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.

Lastly, in selection step, the target vector's lowest value is taken into account. a_i, G_p and the test vector $v_i, G_p + 1$, as shown in Eq. (6).

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

Up until the halting condition is fulfilled, the preceding stages are repeated. Although DE is simple to use and has a wide range of parameter choices, it cannot guarantee the global solution since local optima might stack, especially in high-dimensional space.

MARINE PREDATOR ALGORITHM (MPA)

The primary driving force for MPA was the Lévy and Brownian motions of ocean predators as well as the best encounters rate policy in the biological relationship between predator and prey. The natural principles that govern the optimal foraging strategy and the ratio of predator-prey encounters in maritime habitats are upheld by MPA. This section examines the creation of the MPA algorithm, a straightforward and effective metaheuristic optimization technique.

2.1 MPA formula

Like most metaheuristics, MPA is population-based, meaning that the first response from the first trial is dispersed randomly across the search space:

$$X0 = Xmin + rand\,(Xmax - Xmin) \quad (7)$$

Where $rand$ is a uniform random vector with a range of 0 to 1, and $Xmin$ and $Xmax$ are the lower and upper bounds for the variables. According to the notion of the survival of the fittest, apex predators in nature are better foragers. Consequently, the fittest solution is determined by selecting the greatest predator in the Elite matrix. The arrays in this matrix are in charge of identifying and tracking the prey by using its locations as a reference.

$$Elite = \begin{bmatrix} x^I_{1,1} & x^I_{1,2} \dots \dots & x^I_{1,d} \\ x^I_{2,1} & x^I_{2,2} \dots \dots & x^I_{2,d} \\ x^I_{n,1} & x^I_{n,2} \dots \dots & x^I_{n,d} \end{bmatrix} \quad (8)$$

The top predator vector, represented by $\overline{X^I}$, is copied n times to create the Elite matrix. There are n search agents and d dimensions, respectively. It's important to keep in mind that search agents can include both victims and predators. because by the time the predator is looking for its prey, the victim has already begun to look for its own food. The Elite will be modified at the end of each repetition if a superior predator replaces the dominating predator.

Predators modify their positions in reaction to prey, which is a distinct matrix with the same dimensions as Elite. In short, the most powerful (predator) develops the Elite from the original Prey produced during initialization. As shown, the Prey is as follows:

$$Prey = \begin{bmatrix} x^I_{1,1} & x^I_{1,2} \dots \dots & x^I_{1,d} \\ x^I_{2,1} & x^I_{2,2} \dots \dots & x^I_{2,d} \\ x^I_{n,1} & x^I_{n,2} \dots \dots & x^I_{n,d} \end{bmatrix}_{n \times d} \quad (9)$$

$X_{i,j}$ represents the $j-th$ dimension of the $i-th$ prey in Eq. (9). It should be mentioned that these two matrices play a key role in the whole optimization process.

3. PROPOSED MPAFS

The following is a list of the issues needed to develop the MPAFS paradigm:

3.1 Coded plan

To encrypt the users, we often use a vector of real integers. As seen in the upper portion of Fig. 1, the vector is utilized for features that map randomly to be in the [0,1] interval. Thus, as illustrated in Fig. 1 lower section, if the component value is equal to or greater than 0.5, it will be substituted with 1 so that the feature is selected; if not, the value approximates 0 and the feature is not selected.

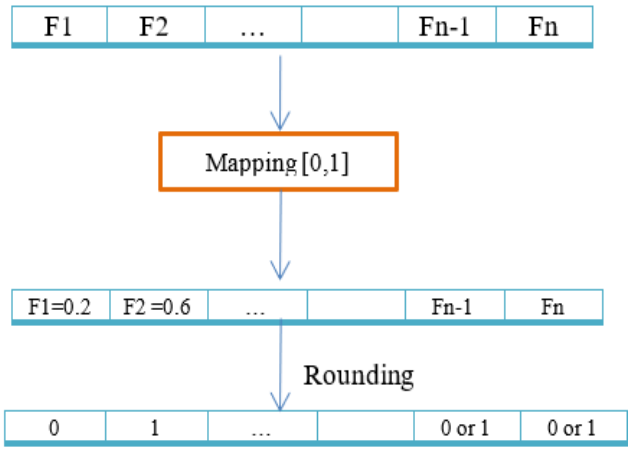

Fig 1: Data encoding and mappings

3.2 Objective function

Our goal function is based on determining the accuracy of each selection; this is done using Eq. (10):

$$Acc = \frac{T_P + T_N}{T_P + F_N + F_P + T_N} \quad (10)$$

Where:

T_P: is the quantity of accurate forecasts, and the actual class is accurate.

T_N: is the quantity of accurate forecasts minus the false actual class.

F_N: is the quantity of accurate predictions compared to the actual class.

F_P: is the quantity of inaccurate forecasts and false actual class.

3.3 System architecture

In this part, we outlined the MPAFS architecture, our suggested system. "System Architecture" was the phrase used in earlier research[16], [17]. The key parts of MPAFS are:

- **Data normalization** is a typical preprocess used in feature selection. In order to prevent the negative effects of some features having bias values, we normalized the features so that they existed in the [0,1] interval. This normalization was accomplished by using FB in Eq. (11) to determine the selected feature.

$$FB = \frac{FA - min_{FA}}{max_{FA} - min_{FA}} \quad (11)$$

- **Individuals decoding:** the chosen features have taken up space in our vector at this stage.

- Selecting training and testing sets: we now separated the dataset into two groups: testing (Xtest, Ytest) and training (Xtrain, Ytrain). The primary features have been represented by X1, X2,... and the primary class is Y, as can be seen in the left portion of Figure 2. We used any classifier, such as SVM, to control Xtrain and Ytrain in order to create the model. To test the model's correctness, we fed it Xtest as input, and we got Y? as the result. If Y? equals Ytest, as seen in the right portion of Fig. 2, ground truth is obtained. Lastly, we use Xtest as the model's input to test the accuracy of the model. The output from the model named Y? is compared to Ytest to see whether they are equal. whether they are, this output. - Select a subset of features: we selected features having a value of 1 from the training set.

- Fitness assessment: we used training set vectors to train our classifier, and we used Eq. 8 to calculate classification accuracy.

- Termination condition: by establishing the maximum number of iterations, we brought the process to an end.

Fig. 3, which illustrates the connections between the key components of the system, provides clarification on the MPAFS system workflow as a whole.

4. RESULTS

MPAFS was used on a portable Intel(R) Core(TM) i7-5500U CPU running at 2.40 GHz with 8 GB of RAM and Windows 10 installed. Matlab R2015a was used to carry out our investigation.

In the current study, we looked at two different types of cancer datasets: synthetic and actual. Tables 1 and 2 list these datasets..

For the years 2010–2012, we used official cancer actual databases for cases of bladder, colon, and breast cancer in Iraqi hospitals. These datasets were tainted by noise, thus bias and irregular values that affected classification performance were eliminated. High performance has been attained using the MPAFS technique using all genuine datasets. The results of MPAFS were compared with those of two other methods, DE-FS and PSO-FS, using datasets from Iraq that included cases of bladder, colon, and breast cancer. The runtime and accuracy are the two parameters used in these comparisons. Results for MPAFS against DEFS and PSOFS algorithms applied to actual datasets for bladder, colon, and breast cancers in Iraq between 2010 and 2012 are listed in Table 3. Figures 4 and 5 display the outcomes stated in Table 3. Results are displayed in Figure 4 with accuracy criteria, and results are browsed in Figure 5 with runtime criteria.

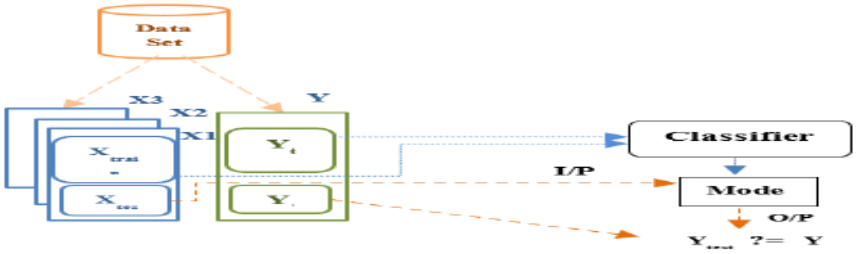

Fig 2: Choosing the procedure for training and testing

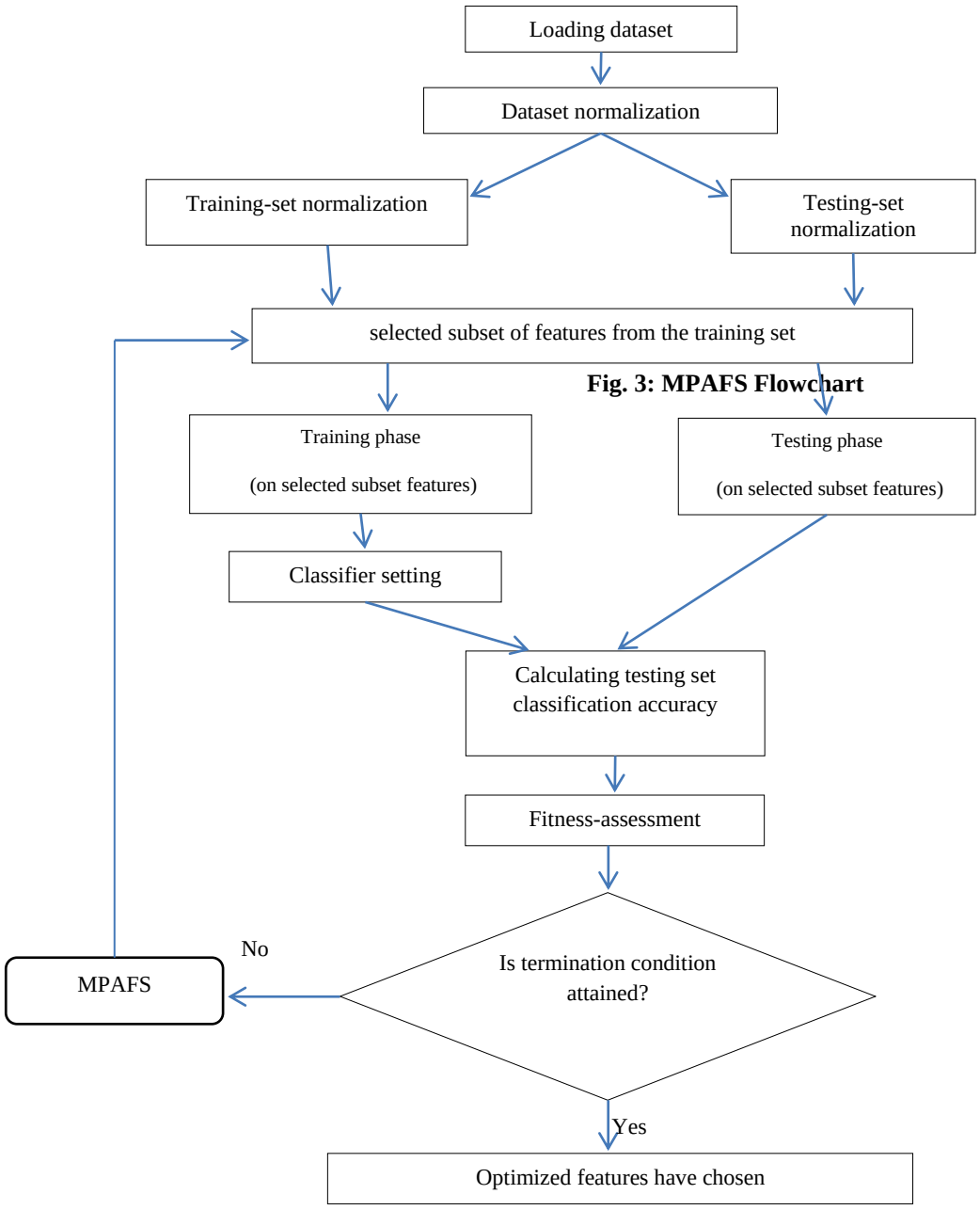

Fig. 3: MPAFS Flowchart

Table 1. Real-world cancer datasets [18]

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 cancer datasets

Datasets	Instances no.	Features no.
Breast Cancer[19]	683	11
Bladder Cancer(Biostat 514/517 Datasets, n.d.)	2922	9
Colon Cancer[20]	1858	16

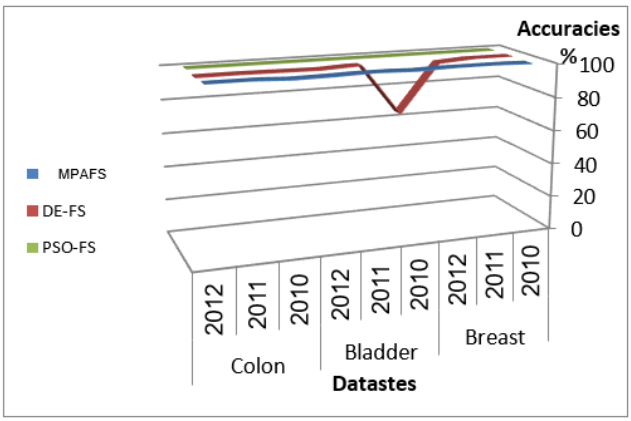

Fig. 4: Findings for MPAFS, DEFS, and PSOFS taking into account accuracy for actual datasets (breast, bladders, and colon in Iraq between 2010 and 2012)

PSOFS has been able to obtain the greatest and closest accuracy levels (99%-100%) using real datasets, as illustrated in figure 4. DEFS, on the other hand, has accuracy values ranging from 70% to 100%.

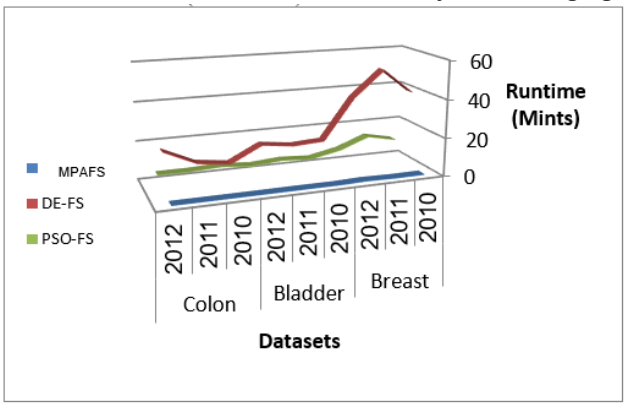

Fig 5: Findings for the MPAFS, DE-FS, and PSO-FS models taking into account runtimes for actual datasets (breast, bladders, and colon from 2010 to 2012 in Iraq)

The authors computed the runtimes using the identical datasets and the three previously described techniques, as shown in figure 5. In this case, we accepted the high convergence rate of MPA to maximize feature selection; for all provided datasets, MPAFS runtimes were under a minute (many seconds). After MPAFS, which takes several minutes to execute, PSOFS comes in second; DEFS, on the other hand, occasionally requires greater runtime, up to an hour. In order to assess our suggested method, MPAFS, we re-ran MPAFS, PSOFS, and DEFS on fictitious datasets pertaining to colon, bladder, and breast cancer. Table 4 presents the results.

Table 3. Outcomes for the MPAFS, DEFS, and PSOFS algorithms using actual datasets for Iraqi breast, bladder, and colon cancer cases between 2010 and 2012.

Dataset	Years	Algorithms	Best Accuracy %	Selected features no.	Run Time (Mints)
Breast-Cancer	2010	MPAFS	99.39	6	0.31
		DEFS	100	8	40.30
		PSOFS	99.93	6	10.17
	2011	MPAFS	100	6	0.32
		DEFS	100	8	52.9
		PSOFS	99.93	7	13.46
	2012	MPAFS	100	6	0.55
		DEFS	98.45	8	38.52
		PSOFS	99.95	6	7.48
Bladder-Cancer	2010	MPAFS	99.34	6	0.08
		DEFS	70.10	9	18.57
		PSOFS	99.92	7	4.39
	2011	MPAFS	100	6	0.26
		DEFS	100	8	17.46
		PSOFS	99.90	8	5.34
	2012	MPAFS	99.34	6	0.10
		DEFS	98.84	9	19.54
		PSOFS	99.89	8	4.1
Colon-Cancer	2010	MPAFS	99	6	0.07
		DEFS	99.06	6	11.54
		PSOFS	99.89	7	5.32
	2011	MPAFS	100	5	0.07
		DEFS	99.26	8	13.60
		PSOFS	99.87	7	4.66
	2012	MPAFS	100	7	0.09
		DEFS	99.15	6	21.02
		PSOFS	99.88	6	4.56

Table 4. Outcomes for the breast, bladder, and colon cancer simulation datasets used in the MPAFS, DEFS, and PSOFS algorithms

Datasets	Algorithms	Best Accuracy %	No. of	Runtime (Mints)
Breast-Cancer	MPAFS	98.77	6	0.04
	DEFS	100	7	6.42
	PSOFS	99.67	8	3.11
Bladder	MPAFS	100	5	0.35
	DEFS	77.02	6	68.41

-Cancer	PSOFS	99.77	7	4.16
Colon-cancer	MPAFS	99.72	5	0.05
	DEFS	66.67	9	38.33
	PSOFS	99.75	8	3.09

Table 4's contents are seen in Figures 6 and 7. Figure 6 depicts the reapplication of the designated methodologies on artificial datasets while taking accuracies into account, whereas Figure 7 takes runtimes into account.

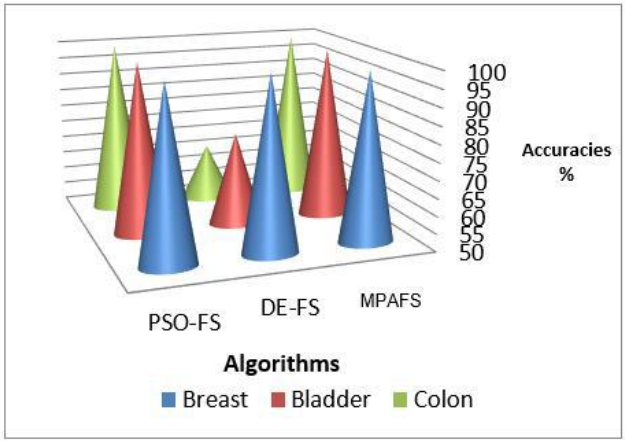

Fig. 6: Outcomes for MPAFS, DE-FS, and PSO-FS taking into account the accuracy of synthetic datasets (colorectal, bladder, and breast cancers)

In line with earlier findings (based on actual datasets), MPAFS and PSOFS achieved the highest accuracy levels, whereas DEFS obtained the lowest accuracy levels.

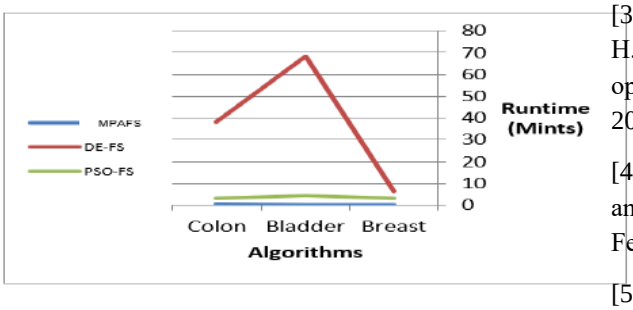

Fig. 7: Outcomes for MPAFS, DE-FS, and PSO-FS taking into account runtimes for synthetic datasets (cancers of the breast, bladder, and colon)

Once more, MPAFS has achieved the lowest runtime (almost nil). The second position was filled by PSOFS, while DEFS, which occasionally took up more than an hour, was last.

3. Conclusion

The Marine Predator technique (MPA), a novel heuristic optimization technique, was developed in this study to present a new feature selection method. The recently suggested method is called MPAFS. Outcomes of this strategy in comparison to earlier strategies like DEFS and PSOFS. The two comparative criteria are runtime and accuracy. We used real and synthetic datasets to apply MPAFS, DEFS, and PSOFS. In comparison to DEFS and PSOFS, we discovered that MPAFS had the best accuracy and the shortest runtimes across all datasets. Due to their propensity to stack in local optima, PSO and DE techniques have a low convergence rate, which is their primary disadvantage. However, we were pleased with the strong rate of convergence for MPA in terms of feature selection across all datasets; this was evident in the short MPAFS runtimes. In addition to the optimization strategy used in the current paper, we suggested using MPA in datamining applications for future work.

5. 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, doi: 10.1016/j.aca.2011.07.027.

[2] C. L. Huang and C. J. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," *Expert Syst Appl*, vol. 31, no. 2, pp. 231–240, 2006, doi: 10.1016/j.eswa.2005.09.024.

[3] 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, doi: 10.1109/SMCIA.2008.5045944.

[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, doi: 10.1016/j.eswa.2011.03.028.

[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, doi: 10.1016/j.swevo.2012.09.003.

[7] O. Ceylan and T. Gulsen, "A Comparison of differential evolution and harmony search methods for svm model selection in hyperspectral image classification CLASSIFICATION O ˘ guzhan Ceylan Kemerburgaz University Department of Electrical and Electronics Engineering Istanbul , Turkey ," *Igarss 2016*, pp. 485–488, 2016.

[8] H. R. Kanan, K. Faez, and S. M. Taheri, "Feature Selection Using Ant Colony Optimization (ACO): A New Method and Comparative Study in the Application of Face Recognition System BT," pp. 63–64, 2007.

[9] H. M. Zawbaa, E. Emary, B. Parv, and M. Sharawi, "Feature selection approach based on moth-flame optimization algorithm," *Proceedings of IEEE Congress on Evolutionary Computation (CEC)*, pp. 4612–4617, 2016, doi: 10.1109/CEC.2016.7744378.

[10] 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.

[11] H. Tariq Ibrahim, W. Jalil Mazher, and E. Mahmood Jassim, "Feature Selection: Binary Harris Hawk Optimizer Based Biomedical Datasets," *Inteligencia Artificial*, vol. 25, no. 70, pp. 33–49, Nov. 2022, doi: 10.4114/intartif.vol25iss70pp33-49.

[12] H. T. Ibrahim, W. J. Mazher, O. N. Ucan, and O. Bayat, "A grasshopper optimizer approach for feature selection and optimizing SVM parameters utilizing real biomedical data sets," *Neural Comput Appl*, 2018, doi: 10.1007/s00521-018-3414-4.

[13] E. Emary, HossamM.Zawbaa, C. Grosan, and A. E. Hassenian, *Feature Subset Selection Approach by Gray-Wolf Optimization*, vol. 334. 2015. doi: 10.1007/978-3-319-13572-4.

[14] A. Faramarzi, M. Heidarinejad, S. Mirjalili, and A. H. Gandomi, "Marine Predators Algorithm: A nature-inspired metaheuristic," *Expert Syst Appl*, vol. 152, Aug. 2020, doi: 10.1016/j.eswa.2020.113377.

[15] R. Storn and K. Price, "Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997, doi: 10.1023/A:1008202821328.

[16] C. L. Huang and C. J. Wang, "A GA-based feature selection and parameters optimizationfor support vector machines," *Expert Syst Appl*, vol. 31, no. 2, pp. 231–240, 2006, doi: 10.1016/j.eswa.2005.09.024.

[17] 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, doi: 10.1007/s00521-016-2818-2.

[18] Ministry of Health-Iraq-Iraqi Cancer Board, *Acceptance of Official Cancer datasets from Iraq*. 2017.

[19] Lichman M, "UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences." Accessed: Jul. 01, 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[20] "Biostat 514/517 Datasets." http://courses.washington.edu/b517/Datasets/datase