Damascus university Journal of Engineering Sciences
42 (No):2 PP: 166-188

مجلة جامعة دمشق للعلوم الهندسية
المجلد 42 (العدد):2 الصفحات: 166-188

A new hybrid model of chaotic Metaheuristic optimizer for feature selection

Mahmod Mohamd Hussien[*1] Ali Salman Barakat[2]
[*1.] Eng, Master of web science in Syrian Virtual University.
mahmodhussien@SyrianVirtualuniversity.edu.sy
[2.] Eng, Master of web science in Syrian Virtual University
alibarakat637@SyrianVirtualUniversity.edu.sy

Received: 31/7/2023
Accepted: 1/10/2023

ABSTRACT:

Recently, there have been many efforts to invent a new ideal model for feature selection in big data applications using multiple approaches including the most important approaches which are those based on Metaheuristics algorithms, in order to solve the problem of the high complexity that faces algorithms applied to big and massive data sets.

Metaheuristics algorithms have made great progress in many fields related to optimization (obtaining the optimal solutions (maximum - minimum)), as they can generate good solutions in reasonable periods of time, but on the other hand they suffer from many disadvantages such as falling into the local optimal solution trap (stuck at local edges), lack of search diversity (occupying the entire search space by search agents) and imbalance between Exploitative and exploratory performance.

In this paper a new hybrid binary feature selection model called “Chaotic Salp Swarm Algorithm Whale Optimization Algorithm” (CSSAWOA) was proposed to solve Multi-objective optimization problems, where the basic idea of CSSAWOA is to improve the Salp Swarm Optimization Algorithm (SSA) by hybridizing it with the Humpback Whale Optimization Algorithm (WOA). In order to share their strengths.

results obtained from applying the proposed model were compared with results obtained from applying some famous and original Metaheuristics algorithms as: Particle Swarm Optimization (PSO), Gray Wolf (GWO), Salp Swarm (SSA), and Humpback Whales (WOA), using /15/ diverse data sets in size, big, medium and small, according to the following criteria: Obtaining the least number of relevant features - Increasing classification accuracy - Reducing processing and computation time.

KEYWORDS: Feature Selection, Metaheuristics, Big Data Analytics, , Salp Swarm Algorithm SSA, Whale Optimization Algorithm WOA.

http://journal.damascusuniversity.edu.sy/index.php/index/index
ISSN (online): 2789-6854

نموذج هجين جديد لمُحسِّن فوضوي ذاتي الإرشاد لاختيار السمات

محمود محمد حسين*[1] علي سلمان بركات[2]

*[1]. مهندس، ماجستير في علوم الويب من الجامعة الافتراضية السورية.
mahmoudhusain60@SyrianVirtualUniversity.edu.sy
[2]. مهندس، ماجستير في علوم الويب من الجامعة الافتراضية السورية.
alibarakat637@SyrianVirtualUniversity.edu.sy

الملخص:

كثرت في الآونة الأخيرة الجهود لبناء نموذج مثالي لاختيار السمات في تطبيقات البيانات الضخمة باستخدام مناهج متعددة، ومن أهمها المناهج المعتمدة على الخوارزميات ذاتية الإرشاد (Metaheuristics)، وذلك من أجل حل مشكلة التعقيد العالي التي تواجه الخوارزميات المطبقة على مجموعات البيانات الضخمة.

حققت الخوارزميات ذاتية الإرشاد (Metaheuristics) تقدماً كبيراً في العديد من المجالات المتعلقة بالتحسين (الاستمثال) أي الحصول على القيم المثلى من النتائج (العظمى – الدنيا)، إذ يمكنها توليد حلول جيدة في فترات زمنية معقولة، إلا أنها من ناحية أخرى تعاني من العديد من العيوب كالوقوع في مصيدة الحل الأمثل المحلي (التعليق عند النهايات المحلية)، والافتقار إلى تنوع البحث (إشغال قطاعات فضاء البحث بالكامل من قبل وكلاء البحث بالشكل الأمثل)، وعدم التوازن بين الأداء الاستغلالي (Exploiting) والاستكشافي (Exploration).

تقترح هذه الورقة البحثية نموذج ثنائي هجين لاختيار السمات يدعى Chaotic Salp Swarm Algorithm Whale Optimization Algorithm (CSSAWOA)، لحل مسائل التحسين متعددة الأهداف، حيث تتمثل الفكرة الأساسية لـ CSSAWOA في تحسين خوارزمية تحسين أسراب السالب (Salp Swarm Algorithm) من خلال تهجينها مع خوارزمية تحسين الحيتان الحدباء (Whale Optimization Algorithm)، بغية مشاركة نقاط القوة فيما بينهما.

تمت مقارنة النتائج التي تم الحصول عليها من تطبيق النموذج المقترح، مع ما تم التوصل إليه من نتائج إثر تطبيق بعض الخوارزميات ذاتية الإرشاد الشهيرة والأصلية وهي: خوارزمية تحسين سرب الجسيمات (PSO)، والذئب الرمادي (GWO)، وأسراب السالب (SSA)، والحيتان الحدباء (WOA)، وذلك باستخدام /15/ مجموعة بيانات متنوعة من حيث الحجم ضخمة ومتوسطة وصغيرة، وذلك وفق المعايير التالية: الحصول على أقل عدد من السمات ذات الصلة – زيادة دقة التصنيف – تقليل زمن المعالجة والحساب.

الكلمات المفتاحية: اختيار السمات، Metaheuristics، تحليلات البيانات الضخمة، خوارزمية تحسين سرب السالب SSA، خوارزمية تحسين الحيتان الحدباء WOA، CSSAWOA.

تاريخ الإيداع: 2023/7/31
تاريخ القبول: 2023/10/1

INTRODUCTION:

High complexity in big data sets known as Curse of dimensionality makes them a major challenge and obstacle specially for analysis, exploration and knowledge extraction, in addition to a number of challenges represented by the increase in the time of building models, redundant and useless information, and the low level of quality - so it was necessary to pay great attention to the process of feature selection, and to make it the most important prior process during processing in order to Reduce the dimensions of the data set, eliminating irrelevant features, choosing the optimal subset of related features from the big data set, increasing the accuracy of the classification and clustering models and patterns which in turn remove noise, extraneous and ambiguous data, and reduce processing and computing time.

Metaheuristics algorithms have made great progress in many fields related to optimization and feature selection, as they can generate good solutions in reasonable periods of time, but on the other hand they suffer from many disadvantages such as falling into the local optimal solution trap, lack of search diversity and imbalance between Exploitative and exploratory performance.

In this paper, a new hybrid Metaheuristics algorithm called CSSAWOA was proposed to solve multi-objective optimization problems. The main idea of CSSAWOA is to improve the Salp Swarm (SSA) algorithm by hybridizing it with the Humpback Whale Optimization (WOA) algorithm.
Salp swarm optimization algorithm has a great ability to exploit, but it suffers from weak exploration because each salp of followers update its position depending on the previous one in the chain, so the proposed model will rely on the strong exploration feature in the whale optimization algorithm, thus the followers will update Their positions based on a random salp position as it used in the whale optimization algorithm in the search for prey phase.
The hybrid design is expected to greatly enhance the exploitation and exploration ability of the proposed algorithm and increase the algorithm convergence speed.

1. RELATED WORK:

Metaheuristics techniques have proven effective in providing solutions through many applications. Such as the feature selection process, which is a multi-objective optimization issue, as it aims to increase classification accuracy, reduce the number of selected features, and reduce computational time.
In this context, the researchers carried out many studies to reach the optimal model that meets the previous requirements.
For example, Sayed et al. (Darwish, Hassanien, & Sayed, 2018) introduced a new Chaotic Whale Optimization Algorithm (CWOA) to solve the problem of falling into the trap of local optima and slow convergence speed, in which chaotic search was integrated with search iterations in the WOA algorithm that uses Ten Chaotic Maps to improve its performance.
The chaos helps the controlling parameter to find the optimal solution more quickly and thus refine the convergence rate of the algorithm.
Although global best solutions are obtained at high speed, scanning the entire search space increases running time of the algorithm. In addition, these modifications make the algorithm lean more towards exploration than exploitation, and thus a balance between them is not achieved.
Chen et al. (Chen, Feng-Yu, & Xian-Feng, 2019) proposed a method based on wrapper technology together with a combination of particle swarm optimization PSO together With SSM algorithm to produce a new algorithm (HPSO-SSM) to obtain the optimal subset of related features. They made three changes to HPSO-SSM:
First, a logistic map sequence is used to enhance the diversity in the search process.
Second, two new parameters are introduced into the original position update formula, which can effectively improve the position quality of the next generation.

http://journal.damascusuniversity.edu.sy/index.php/index/index
ISSN (online): 2789-6854

Finally, a spiral-shaped mechanism is adopted as a local search operator around the known optimal solution region.

because of using logarithmic spiral path, these modifications lead to cover large uncertain search space with less computational time to explore possible solution or to converse particle toward optimum value. in addition, most popular PSO algorithms have the ability to attain near optimal solution avoiding local solution.

Since the Humpback Whale Optimization Algorithm (WOA) involves the spiral path technique, it is assumed that the proposed model will reduce the computational time.

Tu et al. (Tu, Chen, & Liu, 2019) also proposed three different strategies to overcome the limitations of the gray wolf algorithm and increase its effectiveness. The Gray Wolf Optimization (GWO) algorithm is an effective swarm intelligence algorithm for kinds of optimization problems and issues. However, GWO tends to fall into the local optimal solution trap when solving large-scale problems. Social hierarchy is one of the main characteristics of GWO that affects search efficiency, so an improved algorithm called HSGWO was proposed.

First: The wolf pack is roughly divided into two categories: dominant wolves and omega wolves.

Second: An elite enhanced learning strategy is implemented for dominant wolves to prevent misleading low-ranking wolves and improve collective efficiency.

Third: Hybrid GWO and Differential Evolution (DE) strategy of omega wolves is implemented to avoid falling into the local optimal solution trap.

In addition, a new one-dimensional, macro-dimensional hybrid selection strategy for Omega Wolves was designed to balance exploration and exploitation during optimization.

Finally, a (random) perturbation coefficient was used to maintain the diversity of the particle initial distribution and further improve exploration.

These modifications increased the efficiency of the Gray Wolves Algorithm (GWO), especially exploration and avoiding falling into the traps of local optimal solutions. Likewise, it is assumed that the proposed model will increase exploration in the Salp Swarm Optimization Algorithm (SSA).

Abdel-Basset et al. (MohamedAbdel-Basset, El-Shahat, El-henawy, Albuquerque, & Mirjalili, 2020) also proposed a new hybridization of the gray wolf optimizer with a two-stage mutation model to correct the choice function selection of features for wrapper-based classification problems. Where the sigmoid function was used to convert the continuous search space into a binary space in order to match the binary nature of the feature selection problem. Whereas, the two-phase mutation enhances the exploit ability of the algorithm. The purpose of the first mutation phase is to reduce the number of selected features while maintaining high classification accuracy, and the purpose of the second mutation phase is to try to add more informative features that increase classification accuracy. Since the mutation phase can take a long time, the biphasic mutation can be performed with a small probability. Wrapper methods can provide high quality solutions, so one of the most popular wrapper methods called k-Nearest Neighbor was used.

The proposed model is supposed to use a transformation function to convert continuous values to binary to suit the problem of selecting features for big data analytics.

fitness value for the initial set of solutions will be calculated and evaluated to reach the best value in order to increase classification accuracy, as the matter is not about adding and removing features, but rather adjusting their values according to the best solution, and thus it will not take as long as in the previous study.

An improved hybrid approach was proposed by Mafarja et al. (Mafarja, et al., 2020) using the gray wolf optimizer (GWO) and the whale optimization algorithm (WOA) to develop a new feature selection method based on the wrapper approach.

The main goal of this technique is to mitigate the drawbacks of both algorithms, including convergence and local optimal solution.

The proposed model will combine the two Salp swarm optimization algorithms (SSA) and the humpback whale optimization algorithm (WOA) to take advantage of the powerful exploration feature in WOA.

2. WORK METHODOLOGY:

In the proposed model, the chain of salps was divided into two groups, so that each group includes half the number of elements in the chain, and in this way the first group plays the role of the leader, while the second group plays the role of the followers.

In addition, chaotic tent map function was applied to modify the value of the control coefficient c2 used in the original salp swarm optimization algorithm when updating the leader’s position (leader group), as one of the main benefits of optimization using chaotic maps is to reach a fast convergence rate and avoid falling into the traps of local optima, i.e., Enhance exploration.

Also, when initializing the set of initial solutions, chaotic tent function was applied to enhance the diversity of the presence of search agents in the search space, to avoid falling into the local optima trap, and to increase the algorithm's convergence speed.

2-1. SALP SWARM OPTIMIZATION ALGORITHM (Fan, Chen, Zhang, & Fang, 2020):

This algorithm was inspired by simulating the behavior of swarms of Salps in the ocean, and this algorithm is considered as one of the effective Metaheuristics algorithms based on swarm intelligence.

According to the behavior of salp organisms, all individuals are linked with each other to form salp chain consisting of the leader and followers, which in turn helps the salp organisms to coordinate movement and hunting effectively. The leader is the salp at the front of the chain and it is responsible for the process of exploring food sources. The leader depends on the position of food source as a target in order to continuously adjust its position, and after the leader changes its position, the followers update their positions according to the leader's position (Jia, Wang, Gad, & Salem, 2023).

2-2 HUMPBACK WHALE OPTIMIZATION ALGORITHM (Fan, Chen, Zhang, & Fang, 2020):

WOA simulates the distinctive hunting mechanism of humpback whales in the sea. When the whales find small fish and krill near the surface of the sea, the whales encircle them in a downward spiral and then hunt them. This mechanism in hunting is called the bubble net attack (Ebrahimabadi & Mansouri, 2022), according to this series of behaviors. The stages of the algorithm can be divided into three phases: searching for prey, encircling the prey, and the method of attacking with a bubble net, where the stage of encircling the prey takes place at first, and then the attack process takes place with the bubble net, which in turn consists of two stages: shrinking encircling mechanism, and spiral update of positions, in fact, these two phases can take place at the same time. As for the prey search phase, which expresses the algorithm’s exploration phase, the algorithm needs to choose random positions to search for prey, by doing an update of the whales’ positions according to a random search agent.

2-3. CHAOTIC MAPS (Hegazy, Makhlouf, & El-Tawel, 2019):

Optimization using chaos technology is one of the newest search algorithms. Where chaotic maps are used to replace random variables with chaotic variables (Sayed, Khoriba, & Haggag, 2018).

The main benefits of COA optimization algorithms are fast convergence rate and avoidance falling into local optima traps. These benefits can be used to improve the performance of swarm algorithms. Below, table (1) shows five well-known chaos maps (Jia, Wang, Gad, & Salem, 2023).

Table (1) Some of the most popular chaotic map functions (Li, et al., 2021)

No.	Name	Definition	Range
1	Logistic	$x_{i+1} = cx_i(1 - x_i), c = 4$	(0,1)
2	Piecewise	$x_{i+1} = \begin{cases} \frac{x_i}{q} & 0 \le x_i < q \\ \frac{x_i - q}{0.5 - q} & q \le x_i < 0.5 \\ \frac{1 - x_i - q}{0.5 - q} & 0.5 \le x_i < 1 - q \\ \frac{1 - x_i}{q} & 1 - q \le x_i < 1 \end{cases}$	(0,1)
3	Singer	$x_{i+1} = \mu\,(7.86x_i - 23.31x_i^2 + 28.75x_i^3 - 13.301875x_i^4)\ \mu = 1.07$	(0,1)
4	Sinusoidal	$x_{i+1} = cx_i^2 \sin(\pi x_i), c = 2.3$	(0,1)
5	Tent	$x_{i+1} = \begin{cases} \frac{x_i}{0.7} & x_i < 0.7 \\ \frac{10}{3}(1 - x_i) & x_i \ge 0.7 \end{cases}$	(0,1)

2-4. THE PROPOSED ALGORITHM:

In this paper, SSA and WOA will be combined to produce a new algorithm called CSSAWOA. This algorithm adopts the basic structure of the Salp Swarm optimization algorithm and is optimized with WOA.

2-4-1. ALGORITHM WORK STEPS:

(1) The initialization process begins, during which a specific set of N random solutions is generated, and then the chaotic tent function is applied to them in order to enhance their diversity and occupy the entire search space, in addition to setting the values of the algorithm's parameters (coefficients), which are: maximum and minimum bounds ub- lb, the maximum number of iterations T, problem's dimension d which represents the number of original features in the data set which in turn is the number of elements of one solution (the search agent). Also, convergence array is needed to store the best fitness values that are reached during iterations, a variable to store the best solution and another one to store the best fitness value that has been reached so far.

(2) During each iteration of this algorithm:

• a transformation process on the set of initial random solutions will be applied to obtain the binary form using sigmoid function as in equations (1), (2).

$$T(\Delta X_t) = \frac{1}{1+e^{-\Delta X_t}} \qquad \textbf{(1)}$$

(Mafarja, et al., 2020)

$$X_{t+1}^d\,(t+1) = \begin{cases} 1 \ \ if\ u < T(\Delta X_{t+1}) \\ 0 \ \ if\ u \ge T(\Delta X_{t+1}) \end{cases} \qquad \textbf{(2)}$$

(Mafarja, et al., 2020)

Where ΔX_t refers the continuous value of the (i, j) element of the search agent at iteration t, which expresses a feature of the data set in the solution, and this value will be transformed to binary form.

and $X_{t+1}^d\,(t+1)$ value can be 0 or 1 by random number u which take its values from the range [0,1] compared to $T(\Delta X_{t+1})$.

• Calculating fitness values for all initial solutions to obtain and save the best value.

• Calculate the value of the coefficient c1 of the SSA algorithm and the coefficient a of the WOA algorithm using equations (3) and (4) respectively:

$$c1 = 2e^{-\left(\frac{4l}{L}\right)^2} \qquad \textbf{(3)}$$

(Fan, Chen, Zhang, & Fang, 2020)
(Saafan & El-Gendy, 2021)

$$a = 2\,(1 - t/L) \qquad \textbf{(4)}$$

(Mafarja, et al., 2020)
(Saafan & El-Gendy, 2021)

Where l is the current iteration, L is the maximum number of iterations, t is the current iteration and L is the maximum iterations.

• For each search agent, do the following:

- Calculate the value of the coefficients A, C of the WOA algorithm according to equations (5), (6).

$$\vec{A} = 2 \cdot \vec{a} \cdot \vec{r} - \vec{a} \qquad \textbf{(5)}$$

(Fan, Chen, Zhang, & Fang, 2020)
(Mafarja, et al., 2020)

$$\vec{C} = 2 \cdot \vec{r} \qquad \textbf{(6)}$$

(Fan, Chen, Zhang, & Fang, 2020)
(Mafarja, et al., 2020)

Where the value of the coefficient $\vec{a}$ decreases from 2 in the first iteration to the value of 0 in the last iteration, and $\vec{r}$ is a random vector that takes values within the range [0-1].

http://journal.damascusuniversity.edu.sy/index.php/index/index
ISSN (online): 2789-6854

- Update the positions of the salp leader group according to equation (7), taking into account the need to modify the value of coefficient c2 using equation (8).

$$x_j^1 = \begin{cases} F_j + c1\left(\left(ub_j - lb_j\right)c2 + lb_j\right) \\ \quad if\ c3 \geq 0\ (7{:}a) \\ F_j - c1\left(\left(ub_j - lb_j\right)c2 + lb_j\right) \\ \quad if\ c3 < 0\ (7{:}b) \end{cases} \quad \mathbf{(7)}$$

(Fan, Chen, Zhang, & Fang, 2020)
(Saafan & El-Gendy, 2021)

where x_j^1 is the position of the first salp (leader) in dimension j, F_j is the position of the food source in dimension j , ub_j denotes the upper bound of dimension j , lb_j denotes the lower bound of dimension j, and c1 is a random number.

Also, each of the parameters c2, c3 are random numbers that are generated and take their values within the range [0,1] (EL-HASNONY, BARAKAT, ELHOSENY, & MOSTAFA, 2020).

In the proposed model, Tent Chaotic Map function is used in order to improve salp swarm algorithm, so according to the corresponding equation (8) shown in the previous table (1), the value of the control coefficient c2 used in SSA is updated.

$$c2 = \begin{cases} \frac{c2}{0.7} & if\ c2 < 0.7 \\ \frac{(\frac{10}{3})}{(1-c2)} & if\ c2 \geq 0.7 \end{cases} \quad \mathbf{(8)}$$

(Hegazy, Makhlouf, & El-Tawel, 2019)
(Li, et al., 2021)

- Update the positions of the followers group using equation (9), depending on the position of a random salp.

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A} \cdot \vec{D} \quad \mathbf{(9)}$$

(Fan, Chen, Zhang, & Fang, 2020)

$$\vec{D} = \left|\vec{C} \cdot \overrightarrow{X_{rand}} - \vec{X}\right| \quad \mathbf{(10)}$$

(Fan, Chen, Zhang, & Fang, 2020)

Where $\overrightarrow{X_{rand}}$ is a randomly selected search agent (random salp).

- Adjust solutions within the search space in the event that one of them is abnormal outside the search space.

• Storing the necessary results reached in this iteration (generation).

(3) Check the end condition and save the final results.

(4) Evaluate the results and calculate classification accuracy using the KNN classifier.

2-4-2. PSEUDO-CODE OF THE PROPOSED MODEL:

Because of the simple structure of the proposed algorithm, its computational complexity can be directly calculated according to the complexity of both SSA and WOA algorithm. Whereas it is known that the computational complexity of both SSA and WOA depends on three basic operations: initialization, computation of the fitness function, and position update of the population (search agents) in addition to the conversion to binary form process. In the initialization phase, N solutions are created, and so the complexity is O(N), and in the conversion phase to the binary form, the complexity is O(NT). The complexity of calculating and evaluating the fitness and updating positions (including outlier location detection) is:

O (TN + TND) in each iteration, where T is the number of iterations and D is the number of independent variables. Therefore, the final computational complexity of the proposed model can be deduced as follows:

O(CSSAWOA)=O(N)+ O(NT)+ O(TN+TND)
=O(N(1+T+T+TD)) =O(N(1+2T+TD))

Figure (1) shows the pseudo-code of the proposed algorithm CSSAWOA.

```
Initiate the Parameters (e.g., the maximum number of iterations L and population size N)
Initialize the population Xi (i=1, 2,...,n) randomly and by using tent chaotic map Eq.(1)(2)(8)
While(t<L) do
        Transform population Xi (i=1, 2,...,n) to binary
        Calculate the fitness values for search agents
        Set X* as the best search agent
        Calculate c1 using Eq. (3)
        Calculate a using Eq.  (4)
        for each search agent do
            Update parameters A, C using Eq. (5)(6)
            If1(i<=N/2) then
                    Coefficients c2 & c3 take values from [0-1]
                    Update c2 using Eq. (8)
                If2(c3>=0.5) then
                    Update the position of the leader by Eq. (7: a)
                Elseif2 c3<0.5) then
                    Update the position of the leader by Eq. (7: b).
                End if2
            Elseif1 (i>N/2) then
                    Select a random solution(salp) from a population
                    Update the position of the follower using Eq. (9)
            End if1
                Check if any solution goes beyond the search space and adjust it
        End for
        t=t+1
    End While
    Return X*
```

Figure (1) Pseudo-code of proposed model

2-4-3. DEVELOPMENT AND OPERATING ENVIRONMENT:

The proposed model was prepared, designed and applied using Python programming language, and the data sets needed to perform the tests were stored in MongoDB database, which is a non-relational database specialized in the field of big data.

(1) first step: including the required libraries to develop and run the algorithm.

(2) second step: connecting to MongoDB database dedicated to dealing with large data sets and storing the required set of data samples required for testing.

(3) third step: preparing the data to suit the study issue, which is represented by converting it into JSON Data format - then dividing the data set into two groups for training and testing with 70% for training and 30% for testing where each group is consisting of two sub group one for features set and the second for the target (class), then the classification accuracy is measured by the KNN classifier using the K-fold technique to get rid of the problem of overfitting, where the value of k = 10. In addition to set parameters' values and other settings to run the algorithm, which is the number of run times numOfRuns=20, the upper bound where ub=1 and the lower bound lb=0, where these values are chosen in accordance with the values of the features in the data set and by experiment. The number of search agents N and the maximum number of iterations T must also be chosen.

(4) fourth step: Initializing the set of initial solutions (population - search agents) with random and chaotic values by Appling the function of the chaotic tent map, where each solution is nothing but a vector of d elements, which expresses the dimension of the study problem and the number of all features in the dataset.

(5) fifth step: After completing the initialization of population, the appropriate transforming operations must be performed on them to take the binary form instead of the continuous one which is not suitable for feature selection problems by applying Sigmoid transformation function, thus at the end, two values that are 1 when feature is selected and 0 when the feature is not selected.

(6) sixth step: Passing study data set, target array, search agents' array (solution), and the rest of the required coefficients and parameters to run the proposed optimization algorithm, in which the following is done:

• In each iteration, the search agents' array (solution) is converted to binary form and the value of fitness is calculated for all search agents, according to the nature of the issue whether it is minimization or maximization, taking into account that this kind of problem is multi-objective and it is intended to reduce the number of features (minimization) and

http://journal.damascusuniversity.edu.sy/index.php/index/index
ISSN (online): 2789-6854

increase the accuracy of classification (maximization) and to achieve these two contradictory goals, equation (11) is applied.

$$\downarrow \text{Fitness} = \alpha\gamma\text{R}(D) + \beta\frac{|R|}{|D|} \quad \textbf{(11)}$$

(EL-HASNONY, BARAKAT, ELHOSENY, & MOSTAFA, 2020)

where $\gamma R(D)$ is the classification error rate obtained by a given classifier, |R| is the number of selected features, and |D| is the number of all features in the original data set,

$\alpha \in [0,1]$ and $\beta = (1 - \alpha)$ are two parameters related to the importance of classification quality and subset length.

• After completing the calculation of the fitness values for all solutions, the best value will be chosen and store it in the convergence array designated to store the best fitness values in each iteration. At the end of this stage, results are obtained.

(7) seventh step: Evaluating the results reached in terms of: the number of selected features - classification accuracy using the KNN classifier and the computation time (execution time).

2-4-4. DATA SETS:

The proposed algorithm has been studied and tested using a set of well-known databases in the field of feature selection issues in order to verify the strength, efficiency, and effectiveness. Table (2) shows the data sets used in the tests.

The proposed algorithm was applied to 15 databases obtained from the UCI machine learning repository and the following website: https://www.openml.org/search .

1 **Table (2) Characterization of the datasets**

NO	Dataset	Instances	Total Features	Classes
1	Breast Cancer	286	9	2
2	Zoo	101	16	7
3	scene	2408	300	2
4	Sonar	208	56	2
5	Lung Cancer	32	56	2
6	diabetic	1151	19	2
7	parkinsons	195	22	2
8	Climate	540	20	2
9	Ionosphere	351	34	2
10	pageblocks	5473	10	5
11	cpu98lrn	95412	480	4
12	Segment	2310	19	7
13	spectEW	267	44	2
14	Vehical	94	18	4
15	lymphography	148	18	4

2-4-5. PARAMETERS SETTING:

To evaluate the effectiveness of the proposed algorithm in comparison with other most popular contemporary algorithms related to feature selection, it was compared with four well-known algorithms, including the two original algorithms that were used in the hybridization process, which are WOA and SSA, in addition to the standard PSO and the standard GWO.

The initial values of the control parameters c1, c2, c3, a, A, and c were chosen and set according to their values in the original algorithms.

The initial values of the lower and upper bounds ub, lb depend on the principle of choosing the feature or not, that is, zero when the feature is not selected, or one when it is selected.

The values of the rest of the parameters were adjusted according to repeated experimentation, knowing that the higher their values, the more accurate the results and the longer the execution time.

In this context, there are special algorithms concerned with choosing the optimal values for operational parameters like Bayesian algorithm.

20 separate execution cycles were implemented for each algorithm with the same set of data, but with different random seeds of initial solutions each time taking into account transforming them from continuous space to binary space to fit feature selection problem, in addition to applying the chaotic tent map function for initializing Solutions. The maximum number of iterations for all tests is 25.

KNN classifier is a popular wrapper for feature selection and a kind of popular supervised learning algorithm which is simple and fast to implement. with a standard K-fold cross-validation where fold=10. Thus, data set is divided into 10 pieces between training set and test set at a ratio of 3:7.

the number of search agents for the algorithm is set to 25. Each particle or search agent expresses a solution as shown in Table (4). Table (3) shows the values of algorithm's parameters.

Table (3) Control and operating parameters

Parameter	Value	Parameter	Value
K-neighbors	10	c1	Decreased from 2 to 0
Iterations	25	a	Decreased from 2 to 0
Independent runs	20	A	[-1,1]
K-fold cross validation	10	C	[0-2]
Search agents	25	c2, c3	[0-1]
Data dimension	d		
Ub	1	Lb	0

Table (4) Swarm Initialization (Initial Solutions) (EL-HASNONY, BARAKAT, ELHOSENY, & MOSTAFA, 2020)

	d_1	d_2	. . .	d_m
X_1	1	0		1
X_2	1	1		0
.				
.				
.				
X_n	0	1		0

2-4-6. RESULTS COMPARISON:

After doing the experiments and recording the results for each of SSA, WOA, PSO, GWO, and the proposed model, the results can be analyzed according to the following:

2-4-6-1. RESULTS OF THE PROPOSED MODEL:

Table (5) shows the results of the proposed model for the average accuracy of the KNN classifier through 20 independent runs, the computation time and the number of selected features after eliminating the redundant features. The results showed that the proposed model gives us good results in the average of the obtained features and an improvement in classification accuracy for all data sets in a reasonable time.

1 **Table (5) Results of the proposed model**

Dataset	Original Accuracy	Average Accuracy	Computational Time	Total Features	Average Selected features
Breast Cancer	66.28	73.26	2.01	9	5.00

Climate	91.36	95.46	3.10	20	5.70
cpu98lrn	48.00	72.57	60.81	480	92.55
diabetic	58.09	65.98	6.47	19	4.75
Ionosphere	79.25	91.32	2.46	34	3.95
Lung Cancer	80.00	90.00	0.79	56	4.15
lymphography	73.33	90.67	1.17	18	5.55
pageblocks	93.50	94.70	8.93	10	1.90
parkinsons	86.44	96.69	1.41	22	5.80
scene	88.67	91.92	118.38	294	101.10
Segment	89.33	93.06	10.93	19	7.75
Sonar	80.00	100.00	0.80	56	2.90
spectEW	75.31	87.78	1.95	44	13.45
Vehical	41.38	58.79	0.95	18	5.00
zoo	35.48	95.16	1.02	15	5.45

About reducing the dimensions of the data set and selecting features, Figures (4-5-6) present the results obtained in comparison with the total number of features in the data sets used in the study according to the proposed model.

about improving classification accuracy, Figure (7) shows a comparison between the classification accuracy of the entire data set and the classification accuracy of the selected subset of data.

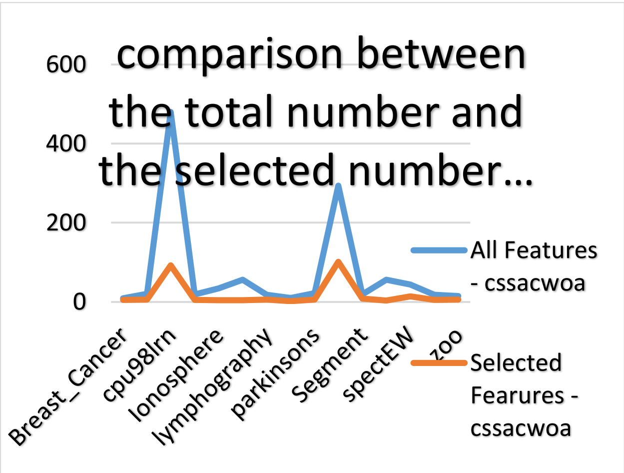


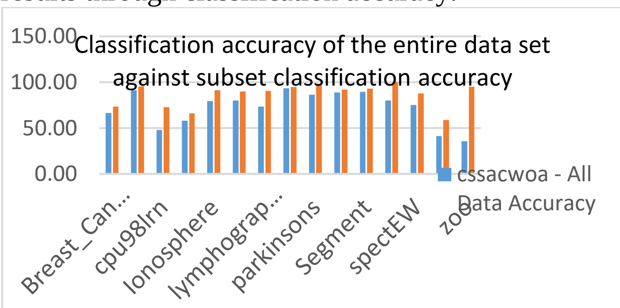
Figure (2) Comparison between total number of features and the average of the selected ones

Figure (3) Comparison of classification accuracy for the total data set and the selected subset

2-4-6-2. NUMBER OF SELECTED FEATURES:

Table (6) presents a comparison between Gray Wolf algorithm (GWO), particle swarm optimization algorithm (PSO), salp swarm optimization algorithm

Figures (2) and (3) presented a comparison between the proposed model and the original data sets with regard to feature selection and classification accuracy, respectively. According to the feature selection issue, the proposed model selects 23.7% of the total features in the entire dataset. Although a small percentage of features were selected, the proposed model achieved excellent and acceptable results through classification accuracy.



SSA, humpback whale optimization algorithm WOA, and the proposed model according to the number of relevant selected features. The results showed the superiority of the proposed model through most data sets and its success in achieving better results, unlike other algorithms, especially with data sets with high dimensions. In Figure (4), a comparison was made with some low-dimensional data sets, while in Figure (5), a comparison with high-dimensional data sets was made, and in Figure (6) a comparison with all data sets.

The proposed model has proven successful in different datasets with different dimensions. Out of the total number of features in all datasets of 1114 features, the proposed model selected 265 features in contrast to the SSA, WOA, GWO and PSO algorithms, which results indicated that the numbers of selected features were approximately as following 531, 655, 457, 499, respectively.

Table (6) shows the number of features selected by the proposed algorithm and the rest of the algorithms compared with the total number of features.

Table (6) Comparison of the average number of selected features by the proposed model and other algorithms

Dataset	Total Features	SSA	WOA	PSO	GWO	CSSAWOA
Breast Cancer	9	3.2	5.65	4.75	4.25	5
Climate	20	11	12	8.65	7.45	5.7
cpu98lrn	480	239.3	284	236	196.45	92.55
diabetic	19	6.4	10.95	6.1	9.6	4.75
Ionosphere	34	17.2	19	13.35	10.55	3.95
Lung Cancer	56	19.6	31.4	13.15	18.9	4.15
lymphography	18	6.8	11.2	6.3	7.95	5.55
pageblocks	10	4.6	5.85	2.75	3.9	1.9
parkinsons	22	7.95	12.1	9.15	8.7	5.8
scene	294	147.45	173.35	141.55	128.2	101.1
Segment	19	10	10.8	10.55	7.55	7.75
Sonar	56	26.7	32.7	15.85	23.7	2.9
spectEW	44	15.3	25.95	21.95	17.1	13.45
Vehical	18	9.85	11.2	4.15	6.75	5
zoo	15	6.15	8.95	5	6.3	5.45
Sum	1114	531	655	499	457	265

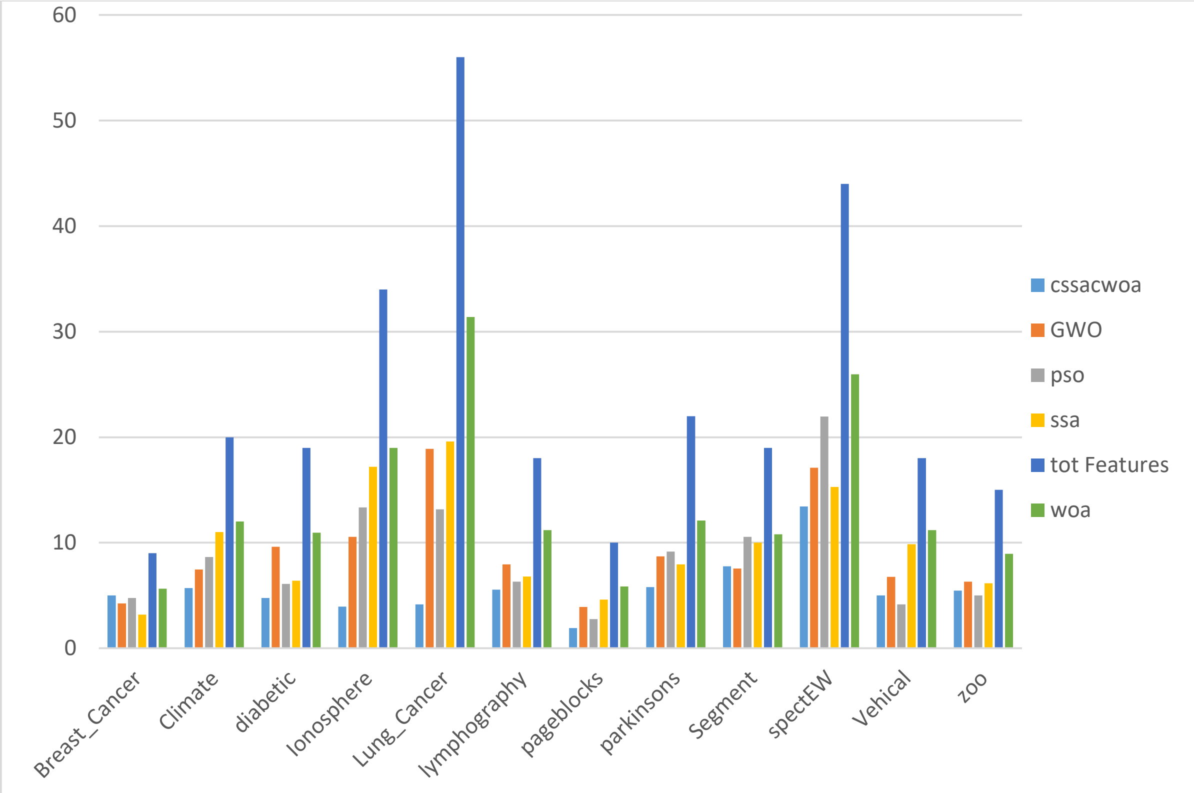


Figure (4) Comparison of the number of selected features with low dimensional data sets

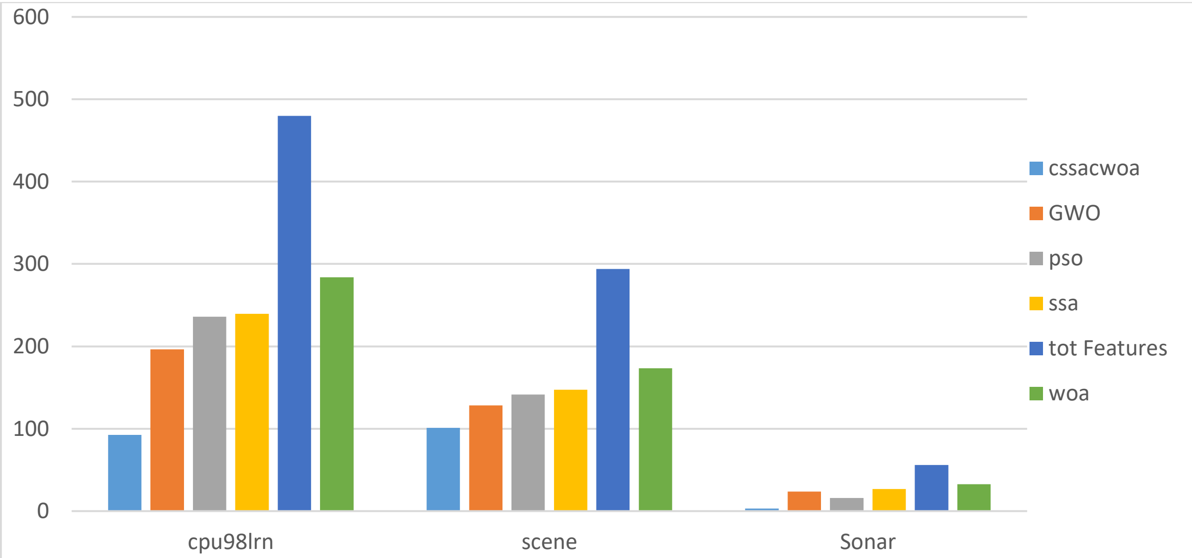


Figure (5) Comparison of the number of selected features with high dimensional data sets

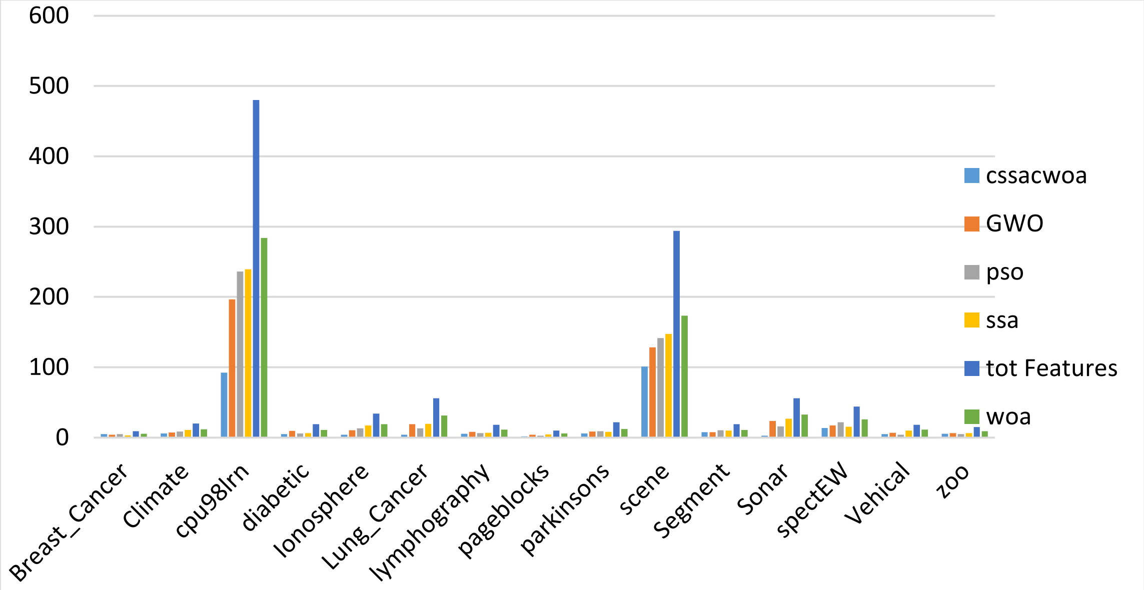


Figure (6) Comparison of the number of selected features with all datasets

2-4-6-3. CLASSIFICATION ACCURACY:

Table No. (7) shows a comparison between the average of classification accuracy for the results of the proposed model with the results of the rest of the algorithms and with the entire data set. As previously mentioned, the KNN classifier was used to calculate the classification accuracy, which is given according to equation (13) (Salgotra, Singh, Singh, Singh, & Mittal, 2021):

$$Accuracy = \frac{Correctly\ classified\ instance}{Total\ number\ of\ instance} \times 100 \qquad (12)$$

The results obtained in most of the data sets confirmed that the proposed model achieved good results in terms of classification accuracy and excellent results when working with big data sets.

In some data sets, the PSO achieved similar results. The results were modeled graphically as shown in Figure (7).

Table (7) Comparison of average of classification accuracy

Dataset	original accuracy	CSSAWOA	SSA	WOA	PSO	GWO
Breast Cancer	66.28	73.26	71.10	68.31	71.80	68.84
Climate	91.36	95.46	93.33	91.57	94.85	91.54
cpu98lrn	48.00	72.57	59.45	49.02	67.95	52.88
diabetic	58.09	65.98	61.97	57.14	65.95	59.26
Ionosphere	79.25	91.32	86.04	79.29	88.73	81.93
Lung Cancer	80.00	90.00	87.00	75.50	80.00	73.50
lymphography	73.33	90.67	88.33	71.44	87.89	85.22
pageblocks	93.50	94.70	94.58	92.62	94.32	93.33
parkinsons	86.44	96.69	94.66	84.32	94.07	88.39
scene	88.67	91.92	90.86	88.80	92.28	89.44
Segment	89.33	93.06	91.81	82.76	92.85	86.48
Sonar	80.00	100.00	96.00	74.00	100.00	84.50
spectEW	75.31	87.78	87.22	74.26	87.04	74.88
Vehical	41.38	58.79	56.72	36.38	56.55	43.79
zoo	35.48	95.16	90.16	57.10	93.55	86.45

Average	72.43	86.49	83.28	72.17	84.52	77.36

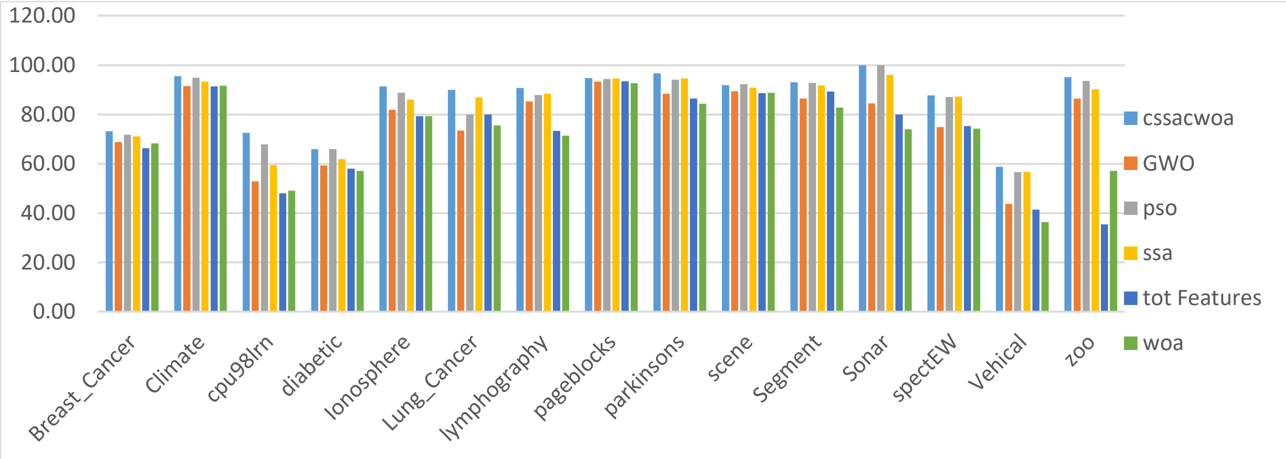


Figure (7) Comparison of classification accuracy

2-4-6-4. COMPUTATION TIME:

A comparison between the proposed model and other related algorithms is made according to the computation time: "The time required for all operations taken by the algorithms is calculated starting from initiating the solutions and ending with obtaining the results and evaluating them after each execution cycle", and at the end the average is obtained by dividing the total time spent by number of runs.

Table No. (8) shows the time spent according to the algorithms used in the experiments, and the results have been illustrated graphically in Figure (8).

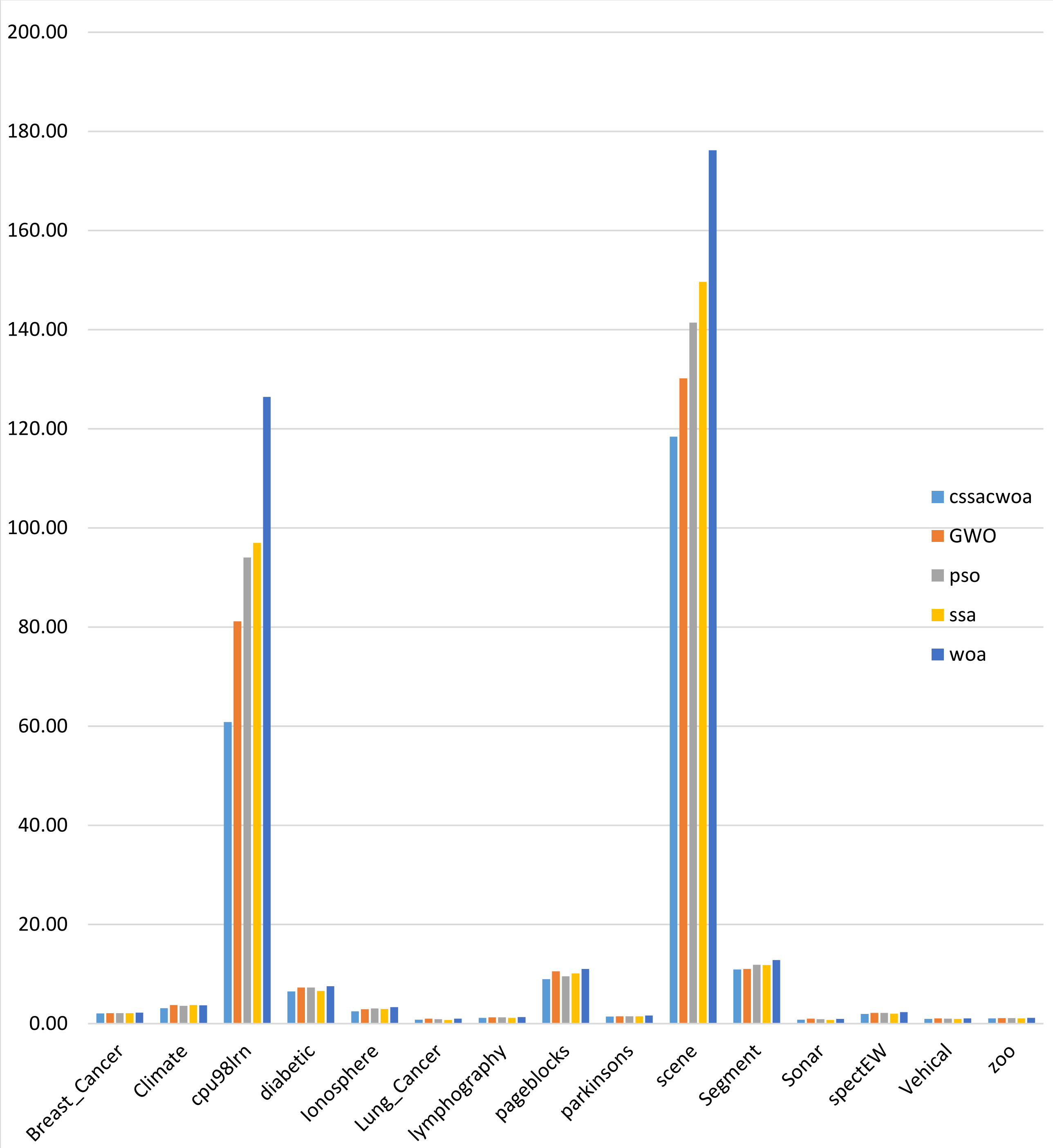
The proposed model achieved less computation time. The total time for all datasets was 221 seconds according to the proposed model, while it equals 292 seconds in SSA, 352 seconds in WOA, 258 seconds in GWO and 282 seconds in PSO.

Noting that with high-dimensional datasets, the model achieved excellent results in terms of time taken to obtain results.

1 **Table (8)Comparison of the computation time**

Dataset	CSSAWOA	SSA	WOA	PSO	GWO
Breast Cancer	2.01	2.09	2.20	2.11	2.09
Climate	3.10	3.73	3.68	3.59	3.72
cpu98lrn	60.81	96.96	126.43	94.03	81.14
diabetic	6.47	6.57	7.52	7.26	7.26
Ionosphere	2.46	2.93	3.29	3.06	2.86
Lung Cancer	0.79	0.75	0.96	0.88	0.98
lymphography	1.17	1.17	1.28	1.23	1.23
pageblocks	8.93	10.10	11.00	9.56	10.52
parkinsons	1.41	1.47	1.61	1.48	1.47
scene	118.38	149.64	176.18	141.39	130.18
Segment	10.93	11.82	12.83	11.88	11.02
Sonar	0.80	0.75	0.96	0.89	1.00
spectEW	1.95	1.99	2.31	2.15	2.12
Vehical	0.95	0.95	1.04	0.98	1.04
zoo	1.02	1.04	1.15	1.06	1.09

sum	221.17	291.95	352.42	281.56	257.73



23

Figure (8) Comparison of the computation time between the proposed model and other algorithms

2-4-6-5. RESULTS' QUALITY:

To prove the quality and effectiveness of the proposed model according to a unified criterion that combines the three previous criteria, which are: the number of selected features - the accuracy of the classification - and the computational time, the following must be followed:

Number of selected features: The minimum number of selected features must be obtained (minimization).

Classification accuracy: The highest possible classification accuracy must be obtained (maximization). To transform the nature of this demand from maximization to minimization, the classification error is adopted, and in this case the classification error must be as low as possible (minimization).

Computational time: The proposed model must also achieve the lowest operating time (minimization).

this problem is one of multi-objective problems, and to achieve harmonization between different objectives, equation (13) is applied, which expresses the quality of the algorithm's result.

This equation was developed by researcher depending on equation (11).

$$\uparrow \text{Quality} = \left(1 - \left(\alpha\gamma R(D) + \beta \frac{|R|}{|D|} + \delta \frac{t}{T}\right)\right) \times 100 \quad (\mathbf{13})$$

Where α, β, δ are weighting factors, and since the previous three criteria are of equal importance, it is necessary that they have the same weighting, thus:

$$\alpha, \beta, \delta = \frac{1}{3} \quad (\mathbf{14})$$

$$\gamma R(D) = \frac{100 - Acc}{100} \quad (\mathbf{15})$$

where $\gamma R(D)$ is the classification error rate obtained by a given classifier, Acc is the classification accuracy of selected subset, |R| is the number of selected features, and |D| is the number of all features in the original data set, t is algorithm's computation time, T is the sum off all algorithms' computation time.

Table No. (9) shows the quality results obtained by the algorithms used in the experiments, and the results are graphically illustrated in Figure (9).

Table (9) Comparison of algorithms' results quality

Dataset	CSSAWOA		SSA	WOA	PSO	GWO	max(quality)
Breast Cancer	66.19		71.88	61.53	66.31	67.24	71.88
Climate	83.19		72.47	70.31	77.15	77.80	83.19
cpu98lrn	80.02		62.83	54.11	66.10	64.76	80.02
diabetic	74.18		69.85	59.36	71.05	62.68	74.18
Ionosphere	87.62		71.79	66.96	76.17	77.10	87.62
Lung Cancer	88.16		78.27	65.80	78.78	72.42	88.16
lymphography	80.20		77.10	62.72	77.55	73.61	80.20
pageblocks	85.96		76.14	70.72	82.58	77.78	85.96
parkinsons	83.79		79.59	69.23	77.53	76.36	83.79
scene	80.33		73.27	68.41	74.79	75.88	80.33
Segment	77.86		72.99	67.99	72.34	75.97	77.86
Sonar	92.21		77.09	64.60	83.82	73.15	92.21
spectEW	79.56		77.84	64.44	72.24	71.95	79.56
Vehical	70.62		60.95	51.06	71.25	61.77	71.25
zoo	79.93		76.59	58.66	80.15	74.70	80.15

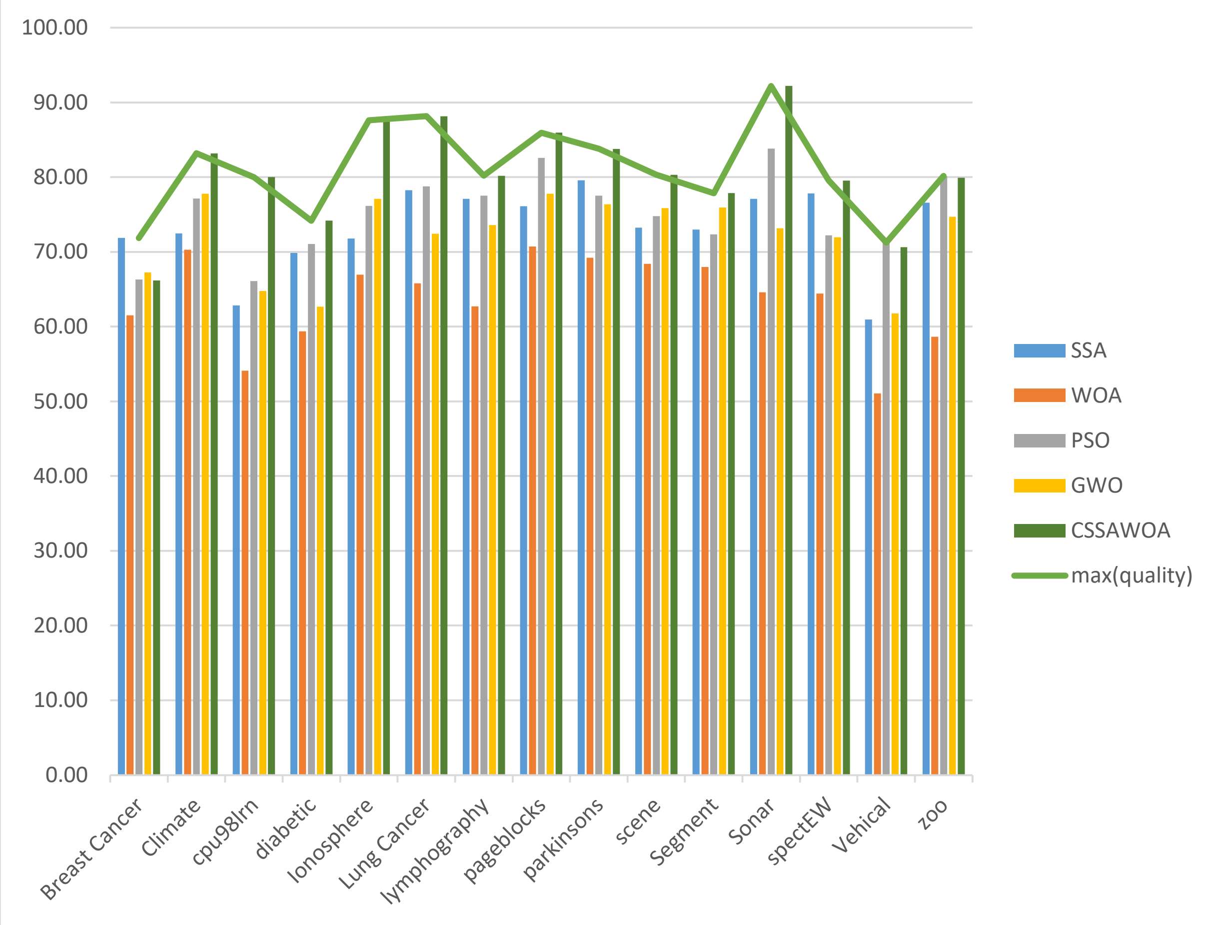


Figure (9) Comparison of algorithms' results quality

Table No. (9) and Figure No. (9) show that the proposed model outperformed the rest of the algorithms in most data sets, while we find that SSA outperformed with "Breast Cancer" data set, and PSO outperformed with "Vehical" and "Zoo" data sets.

2-4-7. RESULTS' DISCUSSION:

In order to prevent the algorithm from falling into local optima trap and to increase the algorithm's convergence speed, the positions of the search agents were initialized with random initial values then to enhance the diversity and quality of them the chaotic tent map function is applied. In addition to modifying the value of the c2 coefficient of the SSA algorithm using the same function (chaotic tent map). In order to overcome the lack of exploration in salp swarm optimization algorithm, the heuristic technique applied in the whale algorithm was followed by updating the position of the salp followers based on the position of a random salp.

In all results, the proposed model achieved better results according to three metrics: classification accuracy, computational time (processing and execution time) and the number of relevant features selected by most of the datasets used.

In a dataset such as Sonar, cpu98lrn and Scene with high dimensionality, the proposed model succeeded in selecting relevant features and discarding most irrelevant features.

Furthermore, in other datasets, there are many achievements regarding classification accuracy, less selected features, and computation time.

In the proposed model, the total number of features selected is very small unlike SSA, WOA, GWO and PSO with reasonable time and good classification accuracy.

According to feature selection issues, the results showed the superiority of the proposed model in all datasets and its success in achieving better results compared to other algorithms.

Out of the total number of features, the proposed model selects 265 features which is equivalent to 23.7% of the total number of features of 1114, in contrast to the SSA, WOA, GWO, and PSO algorithms where the numbers of selected features are approximately 531, 655, 457, and 499, respectively.

As for the classification accuracy of the KNN classifier, the results obtained in most of the datasets confirmed the superiority of the proposed model.

Where the classification accuracy average in all datasets of the proposed model was 86.5%, in contrast, it was 77.4% in GWO, 84.5% in PSO, 83.2% in SSA, and 72% in WOA, while it reached 72.4% for all features in All datasets.

Finally, the proposed model achieves a shorter computation time (execution time).

The total time for all datasets was 221 seconds in the proposed model, while it equals 352 sec in WOA, 292 sec in SSA, 282 sec in PSO, and 258 sec in GWO.

But at the end, to solve all optimization problems, there is no single optimization algorithm available. It is possible that an algorithm's excellent performance with certain data sets may not be so good with other data sets. In some of the data sets used, there may be a decrease in the accuracy rate of the proposed model.

2-4-8. CONCLUSION AND FUTURE WORKS:

This paper proposes a new model to solve and reduce the obstacles when selecting features by using hybridization between SSA and WOA.

The proposed model uses k-nearest neighbor classifier to measure classification accuracy for the subset to find the optimal solutions. problem of falling into local optima trap was solved by random and chaotic initialization using the chaotic tent map function – as well as updating the value of c2 coefficient in the SSA algorithm using the chaotic tent map function. in addition to including the exploration technique used in the whale algorithm by updating the position of the followers based on the position of a random salp.

The transformation function (sigmoid function) was applied to convert the search space from continuous form (continuous vector) to binary form (binary vector) to fit feature selection problem.

To get rid of the overfitting problem, k-fold cross-validation technique was applied.

Many comparisons have been applied with popular and well-known algorithms in this field such as gray wolf algorithm and particle swarm optimization algorithm as well as the original algorithms which are the salp swarm algorithm and humpback whale optimization algorithm.

Fifteen /15/ databases were used to conduct experiments and statistical analyzes to prove the effectiveness and efficiency of the proposed model with all comparison measures.

2-4-8-1. FUTURE WORK:

In the future, the proposed model can be applied to a wide variety of data sets using other transformation functions. and other random and chaotic functions. Moreover, hybridization can be applied between modern Metaheuristic algorithms such as HOA, RFSO, Dragonfly DF and others.

Funding: This Research is funded by Damascus University-Funder no (501100020595).

REFERENCES:

Chen, K., Feng-Yu, Z., & Xian-Feng, Y. (2019). Hybrid particle swarm optimization with spiral-shaped mechanism for feature selection. Expert Systems with Applications, 128, 140-156. doi:https://doi.org/10.1016/j.eswa.2019.03.039

Darwish, A., Hassanien, A. E., & Sayed, G. I. (2018). A New Chaotic Whale Optimization Algorithm for Features Selection. Journal of Classification, 35, 300 - 344. doi: https://doi.org/10.1007/s00357-018-9261-2

Ebrahimabadi, M. A., & Mansouri, N. (2022). A Comprehensive Survey of Feature Selection Techniques based on Whale Optimization Algorithm. doi:https://doi.org/10.21203/rs.3.rs-2388306/v1

EL-HASNONY, I. M., BARAKAT, S. I., ELHOSENY, M., & MOSTAFA, R. R. (2020). Improved Feature Selection Model for Big Data Analytics. IEEE Access, 8. doi:10.1109/ACCESS.2020.2986232

Fan, Q., Chen, Z., Zhang, W., & Fang, X. (2020). Enhanced Whale Optimization Algorithm integrated with Salp Swarm Algorithm for global optimization. Engineering with Computers, 38. doi:https://doi.org/10.1007/s00366-020-01189-3

Hegazy, A. E., Makhlouf, M. A., & El-Tawel, G. S. (2019). Feature Selection Using Chaotic Salp Swarm Algorithm for Data Classification. Arabian Journal for Science and Engineering, 44, 3801–3816. doi:https://doi.org/10.1007/s13369-018-3680-6

Jia, L., Wang, T., Gad, A. G., & Salem, A. (2023). A weighted-sum chaotic sparrow search algorithm for interdisciplinary feature selection and data classification. Scientific Reports, 13. doi:https://doi.org/10.1038/s41598-023-38252-0

Li, J., Guo, L., Li, Y., Liu, C., Wang, L., & Hu, H. (2021). Enhancing Whale Optimization Algorithm with Chaotic Theory for Permutation Flow Shop Scheduling Problem. International Journal of Computational Intelligence Systems, 14(1), 651–675. doi:https://doi.org/10.2991/ijcis.d.210112.002

Mafarja, M., Qasem, A., Heidari, A. A., Aljarah, I., Faris, H., & Mirjalili, S. (2020). Efficient Hybrid Nature-Inspired Binary Optimizers for Feature Selection. journal of Cognitive Computation-SpringerLink, 12, 150-175. doi:https://doi.org/10.1007/s12559-019-09668-6

MohamedAbdel-Basset, El-Shahat, D., El-henawy, I., Albuquerque, V. H., & Mirjalili, S. (2020). A new fusion of grey wolf optimizer algorithm with a two-phase mutation for feature selection. Expert Systems with Applications, 139. doi:https://doi.org/10.1016/j.eswa.2019.112824

Saafan , M. M., & El-Gendy, E. M. (2021). IWOSSA: An improved whale optimization salp swarm algorithm for solving optimization problems. Expert Systems With Applications, 176. doi:https://doi.org/10.1016/j.eswa.2021.114901

Salgotra, R., Singh, U., Singh, S., Singh, G., & Mittal, N. (2021). Self-adaptive salp swarm algorithm for engineering optimization problems. Applied Mathematical Modelling, 89(1), 188-207. doi:https://doi.org/10.1016/j.apm.2020.08.014

Sayed, G. I., Khoriba, G., & Haggag, M. H. (2018). A novel chaotic salp swarm algorithm for global optimization and feature selection. journal of Applied Intelligence, 48, 462–

ISSN (online): 2789-6854

3481. doi:https://doi.org/10.1007/s10489-018-1158-6

Tu, Q., Chen, X., & Liu, X. (2019). Hierarchy strengthened grey wolf optimizer for numerical optimization and feature selection. IEEE Access, 7, 78012 - 78028. doi:10.1109/ACCESS.2019.2921793