

RESEARCH ARTICLE

BinCOA: An Efficient Binary Crayfish Optimization Algorithm for Feature Selection

NABILA H. SHIKOUN^{1,2}, AHMED SALEM AL-ERAQI³, AND ISLAM S. FATHI¹¹Department of Information Systems, Al Alson Higher Institute, Cairo 11762, Egypt²Department of Systems and Computer Engineering, Faculty of Engineering, Al-Azhar University, Cairo 11884, Egypt³Department of Computer Science and Engineering, Faculty of Engineering, Aden University, Aden, Yemen

Corresponding author: Ahmed Salem Al-Eraqi (ahmed_salem7719@yahoo.com)

ABSTRACT The increased utilization of digital instruments like smartphones, Internet of Things (IoT) sensors, cameras, and microphones has resulted in extensive amounts of big data. Inherent challenges associated with big data include significant data dimensionality, redundancy, and irrelevant information. The main objective of feature selection is eliminating unnecessary features, thereby minimizing time and space requirements. This paper proposes a new Binary Crayfish Optimization Algorithm (BinCOA) for feature selection. The Crayfish Optimization Algorithm (COA) is a new metaheuristic algorithm inspired by the simulation of Crayfish search for food, summer resorts, and competitive habits. The original COA has been augmented with two primary enhancements to improve its performance. The refracted opposition-based learning strategy is a novel enhancement incorporated into the initialization step of the COA algorithm to strengthen the algorithm's capability for exploitation. The crisscross strategy is added to the original COA, increasing the COA's convergence accuracy. The algorithm's performance is assessed by evaluating a set of 30 benchmark datasets. The proposed BinCOA is evaluated in comparison to seven contemporary wrapper feature selection methods. The experimental finding indicates that BinCOA consistently outperforms existing algorithms in classification accuracy, average fitness value, and the number of selected features. Furthermore, the statistical significance of the results is verified by calculating the Wilcoxon rank-sum test.


INDEX TERMS Crayfish optimization algorithm (COA), feature selection (FS), classification, the refracted opposition-based learning strategy, the crisscross strategy.

I. INTRODUCTION

Due to the swift adoption of the internet and computer technologies, vast amounts of data, each comprising hundreds of features, are generated. In data mining, the task is to extract valuable information from this extensive dataset to make informed decisions. Meticulously choosing pertinent and beneficial features can significantly influence various applications such as data mining [1], the Internet of Things [2], machine learning [3], and image processing [4]. For instance, within machine learning, redundant, irrelevant, and chaotic records in high-dimensional datasets diminish classification accuracy and escalate computational costs [5]. Keeping and

processing the enormous amounts of data sensors collect is a common problem with IoT techniques. The additional challenge pertains to the existence of irrelevant and redundant features. Consequently, Preprocessing, such as feature selection, is required to handle high-dimensional data and remove redundant or duplicate features. [6]. Feature selection is a crucial aspect of data preparation, playing a significant role in building robust models. It entails identifying and finding the most significant features from the given dataset.

A feature selection framework comprises three primary components: (i) Classification methods like support vector machines (SVMs) [7], k nearest neighbour (kNN) [8], etc., (ii) evaluation criteria, and (iii) the search algorithm employed to identify the most optimal features. Feature

The associate editor coordinating the review of this manuscript and approving it for publication was Sawyer Duane Campbell .

selection methods can be classified into two primary categories: wrapper strategies and filter strategies. Wrapper methods judge feature subsets based on how well they work with the classification algorithm. A wrapper employs the classification algorithm independently, allowing for the assessment of the selected subset's quality depending on its classification effectiveness [9]. A filter approach operates without dependence on any learning model; evaluating subsets of features solely depends on the data, independent of the specific model in use. It's crucial to highlight that filter approaches may not always identify the optimal subset of features. Nonetheless, there's a general observation that wrapper approaches often yield the most optimal feature subset in terms of performance for a predetermined classifier [10]. A feature selection technique aims to pinpoint the optimal subset of features from the entire set of possible subsets. Accurate search methods and metaheuristics are two main types of search algorithms [11]. Accurate search methods explore the entirety of the search space, which, for example, in a feature set with k features, has a magnitude directly related to 2^k , requiring substantial computational resources. Metaheuristic algorithms, on the other hand, exhibit a stochastic nature by initiating their optimization process with randomly generated solutions, effectively exploring the search space. The effectiveness of metaheuristics in addressing feature selection problems relies on their potential to provide solutions approaching optimality within a reasonable timeframe [12]. Due to their simplicity and ease of implementation, metaheuristics display significant adaptability when applied to specific problem domains. A notable feature of these algorithms is their remarkable ability to prevent premature convergence, maintaining a delicate balance between exploration and exploitation, two critical facets.

The Crayfish optimization algorithm (COA) [13] is a newly devised metaheuristic algorithm inspired by the simulation of Crayfish searching for food, summer resort, and competitive habits. The searching for food stage and competitive habits stage represent the exploitation phase of COA, while the summer resort stage constitutes the exploration phase of COA. COA introduces several variables to govern the algorithm's exploration and exploitation, enhancing randomness and optimizing its effectiveness. As a result, we have been prompted to utilize a binary version of COA. As mentioned earlier, metaheuristic algorithms have significantly impacted feature selection issues in the last few years. Despite the extensive research in this field, many metaheuristic algorithms still encounter challenges that require attention. Continued development of optimization techniques is necessary to achieve further improvements in results. So, two primary enhancements have been incorporated into the original Crayfish optimization algorithm (COA) to strengthen its performance. These enhancements reinforce opposition-based learning and cross-cross strategy. The refracted opposition-based learning strategy is implemented to enhance the diversity of the population and minimize the

risk of the method becoming stuck in a suboptimal local state. The crisscross strategy is implemented to strengthen the accuracy of convergence. Our contributions can be summarized in the following points:

- Combining the refracted opposition-based learning strategy with COA, which has the potential to augment the diversity and traversal of the initial population.
- The crisscross strategy is implemented to improve the COA's convergence accuracy.
- BinCOA: A binary modification version of the COA algorithm is proposed to address challenges associated with feature selection.
- The algorithm's effectiveness is assessed through experiments conducted on a collection of 30 well-established benchmark datasets.

The remaining manuscript is structured as follows: Section II presents the literature review, and Section III briefly reviews the Crayfish Optimization Algorithm (COA). The proposed BinCOA algorithm is introduced in Section IV, The Experiments and Analysis are introduced in detail in Section V, and finally, the conclusions are detailed in Section VI.

II. LITERATURE REVIEW

Metaheuristic approaches are commonly categorized into four distinct groups according to the sources that inspire them: human-based methods [14], swarm intelligence [15], evolutionary algorithms [16], and physics-based methods [17]. Human-based methods draw inspiration from how people interact and connect in society. Agrawal [18] introduced a binary variant of the knowledge-based gaining sharing method (GSK) to address feature selection issues, known as FSNBGSK. This approach utilized the k -nearest neighbors (kNN) classifier to assess its performance across 23 benchmark datasets. The proposed method exhibited superior classification accuracy and a minimal number of selected characteristics compared to other algorithms. Examples of algorithms based on human approaches include imperial competition algorithms (ICA) [19], the cultural evolution algorithm (CEA) [20], the volleyball premier league (VPL) [21], and teaching-learning-based optimization (TLBO) [14]. Hybridizing multiple algorithms has become a popular approach in feature selection, enabling researchers to leverage the unique strengths of various algorithms [17]. Swarm intelligence approaches draw inspiration from the collective behaviour of animals in swarms, offering valuable contributions to solving feature selection (FS) problems. Notable algorithms in this category include Binary Horse Herd Optimization (BinHOA) [22], Binary Cuckoo Search (BCS) [23], Binary Dragonfly algorithm (BDA) [24], and Binary Flower Pollination Algorithm (BFPA) [25]. Xue et al. introduced a new approach for Particle Swarm Optimization (PSO) to reduce computational time, minimize the number of features, and maximize the accuracy of classification [26]. Additionally, Al-Tashi et al. [27] presented a binary version

of hybridization modes based on WOA. The SA algorithm is integrated into the WOA framework in the first model. In contrast, the SA algorithm enhances the optimal solution obtained after each iteration in the subsequent model. The results indicate that the methods described in this research are better than existing binary algorithms in terms of accuracy and computational time, with experiments conducted on 18 UCI benchmark datasets.

Evolutionary algorithms emulate the principles of natural evolution, drawing inspiration from the Darwinian theory of evolution. Among these is the genetic algorithm (GA), a type of evolutionary approach for its exceptional ability to effectively address challenges associated with feature selection [28]. The results from implementing nested GA have shown a notable enhancement in classification accuracy. For instance, using the Genetic Algorithm (GA) algorithm in conjunction with chaotic optimization has demonstrated effectiveness in text categorization [29]. Other types of evolutionary methods include differential evolution algorithms (DE) [30], geography-based optimizers [31], and stochastic fractal search [32].

Physics-based methods are formulated from the fundamental principles and rules of natural physics, and metaheuristic algorithms of this nature have significantly contributed to addressing challenges in feature selection. Notable algorithms in this category include the Lightning Search Algorithm (LSA) [33], Multi-verse Optimizer (MVO) [34], Electromagnetic Field Optimization (EFO) [35], Henry Gas Solubility Optimization (HGSO) [36], and Gravitational Search Algorithm (GSA) [37]. Additionally, Simulated Annealing (SA) [38], inspired by metallurgical processes involving controlled heating and subsequent cooling of materials, is considered. The Equilibrium Optimizer (EO) algorithm has emerged as a prominent addition to physics-based approaches in recent years [39]. Ahmed et al. [40] developed an upgraded version of the Equilibrium Optimizer to address feature selection problems. The method was tested on 18 kNN datasets and compared to 8 established approaches, encompassing classical and mixed metaheuristic algorithms. Another development is the binary version of the Equilibrium Optimizer, denoted as BinEO, introduced by D. A. Elmanakhly et al. [3]. This variant incorporates an opposition-based learning method and a local search algorithm [3]. The k-nearest neighbour and SVM classifiers were widely employed as wrapper techniques. A comparative analysis using various established algorithms demonstrated the Binary Equilibrium Optimizer's (BinEO) effectiveness.

The Crayfish Optimization Algorithm (COA) is a novel meta-heuristic algorithm that belongs to the swarm intelligence meta-heuristics algorithms. It mimics crayfish behaviour in competition, summer resorts, and foraging. COA has demonstrated superior performance compared to other widely recognized metaheuristics, showcasing its robust exploration and exploitation capabilities and effectiveness. In our proposed paper, we present utilizing a

binary version of COA as a wrapper feature selection technique to enhance the efficacy of feature selection and classification tasks. To strengthen the performance of COA, the refracted opposition-based learning and cross-cross strategy was combined with the original COA. The refracted opposition-based learning strategy is implemented to increase population diversity and lower the likelihood of the method being caught in an ideal local state. The criss-cross technique is employed to improve the accuracy of convergence.

III. CRAYFISH OPTIMIZATION ALGORITHM

The crayfish is an omnivorous creature that has the ability to consume a wide range of food sources [41]. When crayfish are hunting, they use their claws to tear up big meat and then send it to their second and third feet to hold on to while they walk. Use your second and third walking feet to hold and nibble on small items. As depicted in Figure 3, crayfish commonly employ rapid hiding or utilize pincers as a defensive mechanism to safeguard themselves against potential theft by other crayfish.

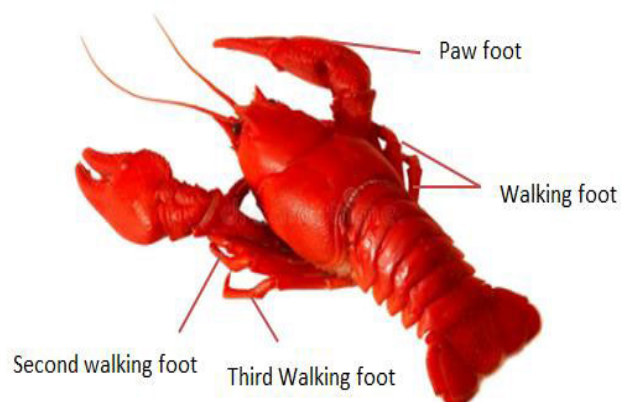


FIGURE 1. The structure of a crayfish.

A. INSPIRATION

COA draws inspiration from the foraging, summer vacation, and competitive behavior of crayfish. The foraging and competition stages might be considered the exploitation stage of the Cultural Evolutionary Approach (COA). In contrast, the summer resort stage can be seen as the exploration stage of COA. The initial part of the algorithm involves defining the crayfish colony Cr to accurately represent the characteristics of optimization by swarm intelligence. The variable Cr_i represents the spatial location of the i th crayfish, serving as an indicator of a potential solution. The regulation of the exploration and exploitation of COA is contingent upon temperature, a stochastic variable that denotes the environmental temperature in which an organism resides. When the ambient temperature exceeds a certain threshold, the COA will transition into the summer resort or competition stage. COA will initiate the foraging stage when the temperature

conditions are suitable. During the foraging stage, the most favorable position for food acquisition is referred to as the optimal solution. The present solution fitness_i (the answer found by Cr_i) and the optimal solution fitness_{food} (the answer found by the optimal solution) both give the size of the food. Crayfish get new positions based on their place when the food is right. Cr_i , food intake stayed the same p , and food placement was Updated on Cr_{foode} . While eating, crabs tear up food with their claw foot if it's too big, then switch between their second and third walking feet to eat.

The sine and cosine formulas were employed to imitate the alternating eating habits of crayfish.

B. MATHEMATICAL FORMULA

1) INITIALIZATIO

The initial step in the process of the Cooperative Optimization Algorithm (COA) involves the generation of a set of Candidate solutions, denoted as Cr , within the given search space. This generation is done randomly. The Candidate solution, denoted a Cr , is formulated with consideration to the population of size N and the dimension (dim). The process of initializing the COA algorithm can be formulated as follows

$$Cr = [Cr_1, Cr_2, \dots, Cr_N] = \begin{bmatrix} Cr_{1,1} & \dots & Cr_{1,j} & \dots & Cr_{1,dim} \\ \vdots & \dots & \vdots & \dots & \vdots \\ Cr_{i,1} & \dots & Cr_{i,j} & \dots & Cr_{i,dim} \\ \vdots & \dots & \vdots & \dots & \vdots \\ \vdots & \dots & \vdots & \dots & \vdots \\ Cr_{N,1} & \dots & Cr_{N,j} & \dots & Cr_{N,dim} \end{bmatrix} \quad (1)$$

where Cr indicates the initial position of the population, N indicates the population's numbers, dim indicates the dimension of the population, $Cr_{i,j}$ is individual positions of i -th in the j -th dimension, and $Cr_{i,j}$ is calculated as follows:

$$Cr_{i,j} = lb_j + (ub_j - lb_j) \times rand \quad (2)$$

where ub_j and lb_j are the upper and lower bounds of the j -th dimension, respectively, and $Rand$ indicates a random number.

2) EFFECT TEMPERATURE ON CRAYFISH INTAKE

Fluctuations in temperature can influence crayfish behavior, prompting transitions between various stages. When the ambient temperature exceeds 30 °C, crayfish exhibit a preference for seeking out cooler environments as a means of engaging in their summer retreat. Crayfish will engage in foraging activity when exposed to suitable temperature conditions. The quantity of food consumed by crayfish is influenced by temperature. The optimal feeding range for crayfish falls from 15°C to 25°C, with 30°C being particularly favorable. Consequently, It is possible to model the feeding quantity of crayfish using a normal distribution.,

illustrating the impact of temperature on their feeding behavior. Due to the robust foraging behavior exhibited by crayfish within the 20 to 30°C temperature range, COA defines a temperature range extending from 20 to 35°C. The equation for temperature is eq.3. The representation of crayfish intake is depicted in eq.4. Figure 2 illustrates the schematic of food intake.

$$Temperature = rand \times 15 + 20 \quad (3)$$

$$p = W \times \left(\frac{1}{\sqrt{2} \times \pi \times \sigma} \times \exp \left(-\frac{(Temperature - \mu)^2}{\sigma^2} \right) \right) \quad (4)$$

where μ stands for the ideal temperature for crayfish, and σ and W are employed to regulate the amount of crayfish consumed at different temperatures.

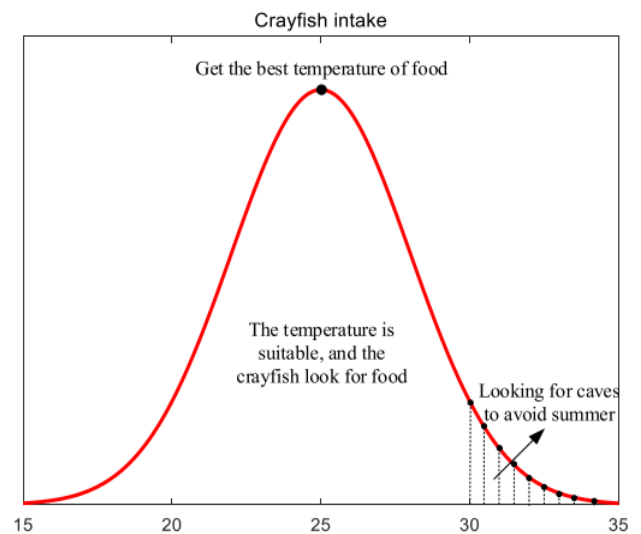


FIGURE 2. The influence of temperature on intake on crayfish [13].

3) PHASE OF SUMMER RESORT (EXPLORATION)

If the temperature is more than 30, it's too hot. At this point, the crayfish decide to spend the summer in the cave. The cave X_{shade} is described as follows:

$$Cr_{shade} = \frac{Cr_G + Cr_L}{2} \quad (5)$$

where Cr_G denotes the optimal position achieved through the cumulative iterations while Cr_L signifies the optimal position within the current population. The phenomenon of crayfish engaging in territorial disputes within caves can be characterized as a stochastic occurrence. When the value of the random variable $rand$ is less than 0.5, it indicates the absence of any rival crawfish for caverns, hence resulting in the direct entry of the crawfish into the cave for the purpose of summer vacation. This process is simulated as follows:

$$Cr_{i,j}^{t+1} = Cr_{i,j}^t + S \times rand \times (Cr_{shade} - Cr_{i,j}^t) \quad (6)$$

In this context, t is the current number of iterations, and $t + 1$ signifies the iteration number for the next generation. Additionally, S is a decreasing curve, as depicted in the following equation:

$$S = 2 - \left(\frac{t}{T}\right) \quad (7)$$

where T is the maximum iteration number.

During the summer resort phase, crayfish aim to approach the cave, symbolizing the best solution. In this phase, crayfish move closer to the cave, effectively bringing individuals nearer to the best solution. This process strengthens COA's exploitation ability, facilitating faster algorithm convergence.

4) PHASE OF COMPETITION (EXPLOITATION)

When the temperature is above 30 degrees, and the $rand$ is less than 0.5, It's a sign that the cave is appealing to more than just crayfish. Eq. 8 shows the crayfish vying for control of the cave.

$$Cr_{i,j}^{t+1} = Cr_{i,j}^t - Cr_{z,j}^t + Cr_{shade} \quad (8)$$

where z is a randomly selected individual of crayfish.

$$z = \text{round}(rand \times (N - 1)) + 1 \quad (9)$$

Crayfish engage in competition with one another, and crayfish Cr_i adjust their positions depending on the position Cr_z of another crayfish. This positional adjustment expands the COA search range, thereby enhancing the algorithm's exploration capability.

5) PHASE OF FORAGING (EXPLOITATION)

When the temperature is equal to or below 30°C, it is considered suitable for crayfish feeding. During this period, Crayfish exhibit active locomotion as they approach the food source. After discovering the food, crayfish evaluate the dimensions of the food item. Crayfish use their claws to dismantle big food items, then ingest them by alternating between their second and third ambulatory appendages.

$$Cr_{food} = Cr_G \quad (10)$$

The size of food Q is presented as:

$$Q = k \times rand \times \left(\frac{fitness_i}{fitness_{food}}\right) \quad (11)$$

where k represents the food factor, signifying the maximum food size with a constant value of 3. $fitness_i$ denotes the i -th crayfish fitness value, while $fitness_{food}$ is the fitness value associated with the location of the food.

The crayfish's assessment of food size is based on the dimensions of the largest food item. When $Q > (k + 1)/2$, it indicates that the portion size of the food is excessive. At present, the crayfish engages in the act of tearing its food

using its foremost claw appendage. The following equation simulates this process:

$$Cr_{food} = \exp\left(-\frac{1}{Q}\right) \times Cr_{food} \quad (12)$$

The equation for foraging, considering the relationship between the food obtained by crayfish and food intake, is as follows:

$$Cr_{i,j}^{t+1} = Cr_{i,j}^t + Cr_{food} \times p \times (\cos(2 \times \pi \times rand) - \sin(2 \times \pi \times rand)) \quad (13)$$

at $Q \leq (k + 1)/2$, The crayfish simply needs to approach the meal and consume it directly. The equation can be expressed as:

$$Cr_{i,j}^{t+1} = (Cr_{i,j}^t - Cr_{food}) \times p + p \times rand \times Cr_{i,j}^t \quad (14)$$

Crayfish employ several feeding methods depending on the size of their food Q , with food Cr_{food} representing the best solution. If the size of the food Q is appropriate for crayfish consumption, the crayfish will approach the food. When the value of Q is excessively large, it signifies the presence of a substantial disparity between the crayfish and the ideal solution. Hence, it is to decrease the prevalence of Cr_{food} and facilitate its proximity to the food. During the foraging step, the COA algorithm will strive to reach the best solution, hence improving its exploitation ability and exhibiting strong convergence capabilities. The flowchart illustrating the process of COA is depicted in Figure 3.

IV. PROPOSED BinCOA

This section thoroughly elucidates the proposed BinCOA, an approach based on wrappers specifically crafted to address the challenge of Feature Selection. The primary steps of the BinCOA algorithm include Initialization using the Refracted Opposition-Based Learning strategy, the transformation function, the Crisscross strategy, and the evaluation. The subsequent subsections will delve into a detailed explanation of each step.

A. INITIALIZATION WITH THE REFRACTED OPPOSITION-BASED LEARNING STRATEGY

Efficient utilization of the local space plays a crucial role in pursuing an optimal solution, significantly impacting the obtained optimal solution quality. Our proposed algorithm, BinCOA, incorporates the Refracted Opposition-Based Learning strategy [42] to enhance the population's initialization. The solution space is expanded by the acquisition of an opposition-based solution derived from the existing solution, hence facilitating the identification of a more optimal alternative solution to address a given problem. The integration of the metaheuristic with opposition-based learning has been established to demonstrate the effective enhancement of algorithmic solution accuracy. In the initialization phase of BinCOA, refracted opposition-based learning is employed to adjust the positions of crayfish within the search space.

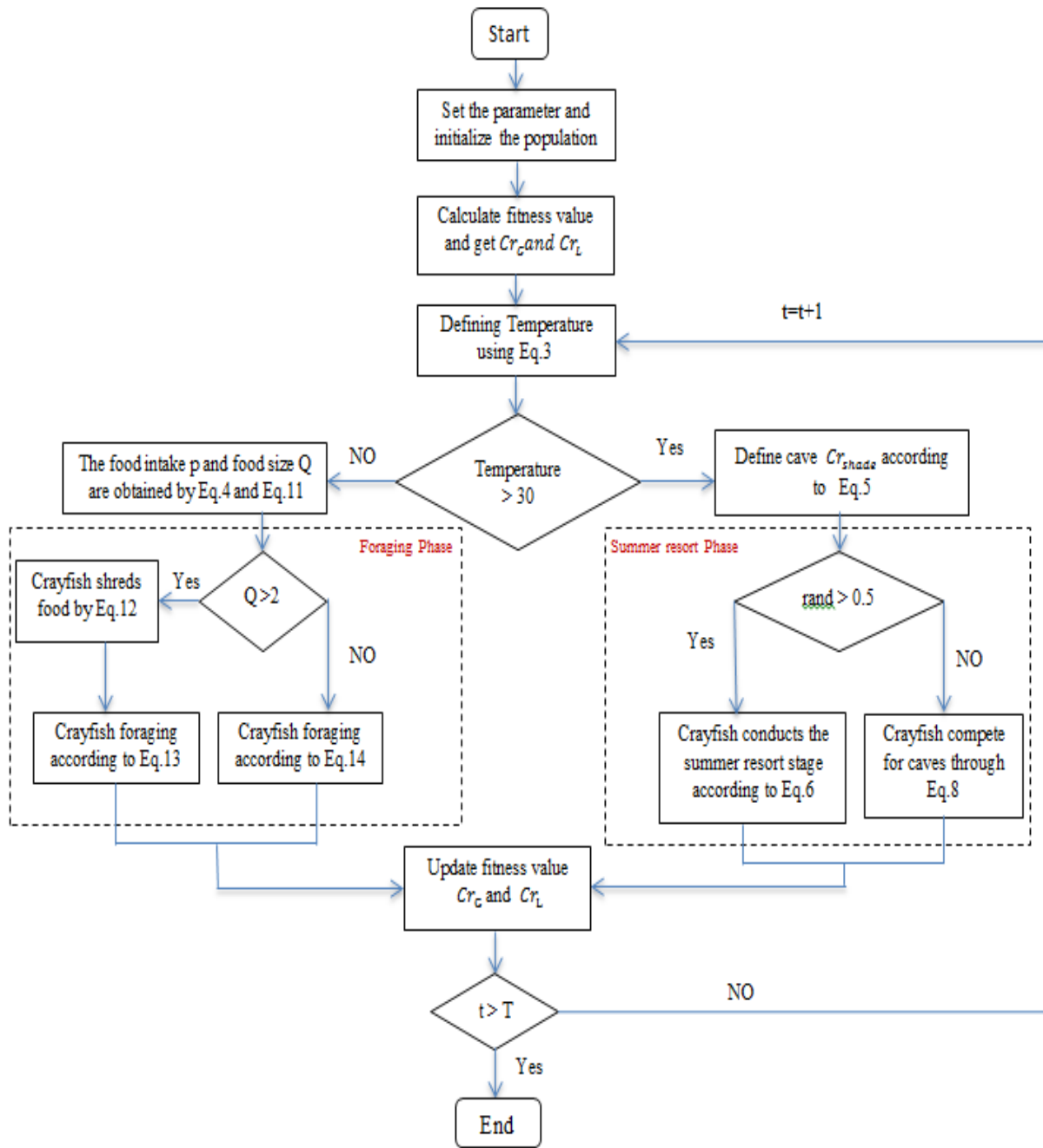


FIGURE 3. The flow chart of crayfish optimization algorithm (COA).

The concept of refracted opposition-based learning is illustrated in Figure 4.

The search interval for solutions on the x-axis extends within the range [lb, ub]; the origin O is situated at the midpoint of the interval [lb, ub]. Here, α and β are assigned as the angle of incidence and the angle of refraction, respectively. Additionally, m and m^* denote the lengths corresponding to the incident and refracted rays, respectively. The refracted formula can be expressed as follows:

Put $\sigma = \frac{m^*}{m}$ and $n=1$ in Eq.15, and COA is extended to a high-dimensional space, resulting in the solution for the refracted direction $Cr_{i,j}^*$, as follows:

$$n = \frac{\sin\alpha}{\sin\beta} = \frac{\frac{lb+ub}{2} - x}{Cr^* - \frac{lb+ub}{2}} \times \frac{m^*}{m} \quad (15)$$

$$Cr_{i,j}^* = \frac{lb_j + ub_j}{2} + \frac{lb_j + ub_j}{2\sigma} - \frac{Cr_{i,j}}{\sigma} \quad (16)$$

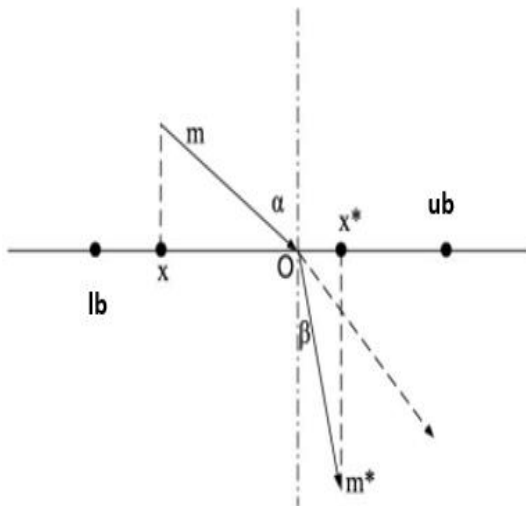


FIGURE 4. The Refracted opposition-based learning.

where $Cr_{i,j}$ is the i -th crayfish position at j -th dimensions, $Cr_{i,j}^*$ is the refracted inverse solution of $Cr_{i,j}$, and lb_j and ub_j are the lower and upper bounds of the dynamic boundary.

B. TRANSFORMATION FUNCTION

The Feature Selection process has traditionally been conceptualized as a binary problem. However, the positions of particles generated by the original Crayfish Optimization Algorithm (COA) are characterized by continuous values. Consequently, to convert the continuous space of the original COA into a binary search space, introducing a transformation function becomes imperative. In the context of feature subset selection challenges, the concentrations of particles are constrained to binary values of 0 or 1. Figure 5 depicts the binary representation of a COA solution designed for a dataset comprising D features. The values of 1 and 0 signify the selected or unselected of the corresponding feature.

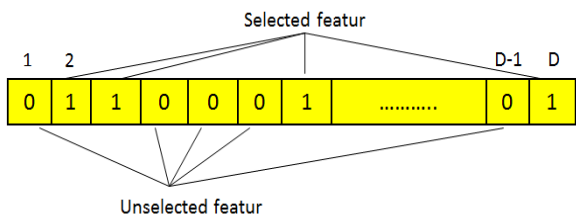


FIGURE 5. Binary representation for BinCOA solution.

The proposed binary COA algorithm employs a binarization technique to transform each solution into its corresponding binary representation. The sigmoid function stands out as one of the most frequently utilized transformation functions within the S-shaped family [43]. The sigmoidal function can be classified as a member of the S-shaped family of transfer

functions, described as follows:

$$T(Cr_i^d(t)) = \frac{1}{1 + e^{-Cr_i^d(t)}} \tag{17}$$

where $Cr_i^d(t)$ the i -th crayfish position. In order to obtain the binary value, the concentration of i -th crayfish is updated according to the following procedure:

$$Cr_i^d(t + 1) = \begin{cases} 1 & rand \geq T(Cr_i^d(t)) \\ 0 & rand < T(Cr_i^d(t)) \end{cases} \tag{18}$$

The variable $rand$ represents a randomly generated value inside the interval $[0,1]$.

C. APPLYING CRISSCROSS STRATEGY TO BinCOA

In this section Crisscross Strategy is described in details to enhance the solution accuracy of the BinCOA algorithm by applying Horizontal crossover and vertical crossover.

1) HORIZONTAL CROSSOVER

The arithmetic crossover applied across all dimensions between two agents is referred to as horizontal crossover [44]. Suppose the i -th crayfish, Cr_i and the k -th crayfish Cr_k are employed to execute the horizontal crossover operation at the j -th dimension. This can be formulated as:

$$Cr'_{i,j} = r_1 \times Cr_{i,j} + (1 - r_1) \times Cr_{k,j} + C_1 \times (Cr_{i,j} - Cr_{k,j}) \tag{19}$$

$$Cr'_{k,j} = r_2 \times Cr_{k,j} + (1 - r_2) \times Cr_{i,j} + C_2 \times (Cr_{k,j} - Cr_{i,j}) \tag{20}$$

where $Cr'_{i,j}$ and $Cr'_{k,j}$ represent the moderation solutions generated as offspring from $Cr_{i,j}$ and $Cr_{k,j}$, respectively. r_1 and r_2 are randomly selected from the range $[0,1]$, while c_1 and c_2 are randomly chosen from the interval $[-1,1]$. To maintain superior crayfish, comparing the solutions generated by the horizontal crossover operation with the pre-crossover solutions is essential.

2) VERTICAL CROSSOVER

Vertical crossover involves applying an arithmetic crossover to all agents between two dimensions [44]. Suppose the i_1 -th and the i_2 -th dimensions of the crayfish Cr_i , they are employed for conducting the vertical crossover operation.

$$Cr'_{i,j} = r \times Cr_{i,j1} + (1 - r) \times Cr_{i,j2} \tag{21}$$

where $Cr'_{i,j}$ is the offspring of $Cr_{i,j1}$ and $Cr_{i,j2}$, r are randomly selected from the range $[0,1]$. The solutions produced through the vertical crossover operation must be compared with the pre-crossover solutions to preserve crayfish better.

D. THE EVALUATION FUNCTION

Choosing a higher number of features from the data presents a challenge, as the classifier's performance tends to degrade when faced with irrelevant or redundant features. Therefore, it becomes crucial to address this issue by reducing the dimensionality of the data. Feature selection emerges as

Algorithm 1 Pseudo-Code of BinCOA

1. Initialization T, Population N, dimension dim
2. Initialize the candidate solutions using Eq. 1 and Eq. 2
3. Apply refracted opposition-based learning using Eq. 11 and Eq.12
4. Evaluate the fitness values of the population to get Cr_G, Cr_L
5. While ($t < T$)
6. Transform the Crayfish positions into binary space by employing a transfer function using Eq.15 and Eq.16.
7. Evaluate each Crayfish within the population by employing kNN or SVM classifiers.
8. Measure the fitness of the entire population of the Crayfish using Eq.20.
9. Defining temperature by Eq.3
10. if (temperature > 30)
11. Define cave Cr_{shade} according to Eq.5
12. if (rand < 0.5)
13. Crayfish conducts the summer resort stage according to Eq.6
14. Else
15. Crayfish compute for caves through Eq.8
16. End if
17. Else
18. The food intake P and food size Q are obtained by Eq.4 and Eq.11
19. if $Q > 2$
20. Crayfish shreds food by Eq.12
21. Crayfish foraging according to Eq.13
22. Else
23. Crayfish foraging according to Eq.14
24. End if
25. Update the position of Crayfish by using the crisscross strategy based on Eq.16 and Eq.18
26. End if
27. Update fitness values, Cr_G, Cr_L
28. $t = t + 1$
29. End While

a technique aimed at improving the efficiency and effectiveness of a given classifier by eliminating unnecessary or irrelevant features. In evaluating solutions, it is not only the classification accuracy rate that is scrutinized; the number of selected features also plays a significant role. In cases where two solutions demonstrate identical classification accuracy, preference is given to the solution with the fewest selected features. Thus, the objective of the fitness function is to optimize the classification accuracy rate by minimizing the classification error while concurrently reducing the number of selected features. The fitness function provided below serves as the metric for evaluating BinCOA solutions, striking a balance between these two primary objectives.

$$fitness = \alpha \gamma + \beta \frac{S}{N} \tag{22}$$

TABLE 1. Description of the datasets.

NO	Dataset	No. of features	No. of instances	No. of classes
1	Breast Cancer	9	699	2
2	Fri_c0_1000_10	10	1000	2
3	Fri_c0_1000_10	10	1000	2
4	Vowel	12	990	11
5	HeartEW	13	270	5
6	Congress	16	434	2
7	Hepatitis	19	155	2
8	Parkinsons	22	195	2
9	IonosphereEW	34	351	2
10	Dermatology	34	366	6
11	WaveForm	40	5000	3
12	Lung Cancer	56	32	3
13	Spambase	57	4601	2
14	SonarEW	60	208	2
15	Splice	60	3190	3
16	Movementlibras	90	360	15
17	Robot1	90	88	4
18	Robot2	90	47	5
19	Hillvalley	100	1212	2
20	Clean1	168	476	2
21	DNA	180	3186	3
22	Semeion	256	1593	10
23	USPS	256	9298	10
24	Arrhythmia	279	452	16
25	LSVT	310	126	2
26	Madelon	500	2600	2
27	CNAE	856	1080	9
28	Colon cancer	2000	62	2
29	Leukemia	7129	72	2
30	Arcene	10000	200	2

TABLE 2. BinCOA parameter configuration.

parameter	value
No. of runs	20
No. of iterations	30
No. of search agents	10
Dimension	No. of features
β	0.01
α	0.99
K-neighbors	5
K-folder cross-validation	10

where $\alpha \in [0, 1]$, γ indicates the classification error rate computed by the kNN or SVM classifier, $\beta = 1 - \alpha$, S represents the selected features, and N is the total features. In the proposed algorithm (BinCOA), kNN or SVM is used as a classifier [7], [45]. We use the SVM classifier method when a dataset has two classes. In every other case, the kNN algorithm is used. The procedural steps for the BinCOA are illustrated in Algorithm 1.

V. EXPERRIMENTAL RESULTS AND ANALYSIS

In this section, we present the outcomes of the suggested methodology and compare them with the latest algorithms. Both the proposed and recent algorithms underwent testing on a laptop with the following specifications: the Matlab R2016a Software operating on the Windows 8 OS, powered

TABLE 3. Results for BinCOA compared to COA in average Fitness, average accuracy, and average No. of selected feature overall datasets.

NO	Dataset	Average Fitness		Accuracy classification		No. of selected feature	
		COA	BinCOA	COA	BinCOA	COA	BinCOA
1	Breast Cancer	0.0312	0.0300	0.8903	0.9819	5	4
2	Fri_c0_1000_10	0.1108	0.1033	0.8261	0.8997	6	5.5
3	Fri_c0_1000_10	0.0955	0.0915	0.9004	0.9314	4	3
4	Vowel	0.0610	0.0590	0.9111	0.9441	8	7.04
5	HeartEW	0.3216	0.3101	0.5741	0.6873	5	5
6	Congress	0.0334	0.0317	0.9204	0.9805	6	6
7	Hepatitis	0.119	0.117	0.8443	0.8911	8	7.35
8	Parkinsons	0.0581	0.0610	0.8961	0.9484	9	7.35
9	IonosphereEW	0.0462	0.0434	0.9141	0.9607	10	11
10	Dermatology	0.0270	0.0219	0.9301	0.9997	14	13
11	WaveForm	0.1365	0.1311	0.8241	0.8771	21	21
12	Lung Cancer	0.2661	0.2443	0.5874	0.6608	16	15
13	Spambase	0.0693	0.0668	0.9314	0.9414	30	28
14	SonarEW	0.1187	0.1003	0.8451	0.8833	22	20.06
15	Splice	0.1733	0.1665	0.8301	0.8378	21	20
16	Movementlibras	0.2501	0.2022	0.7339	0.7934	33	30
17	Robot1	0.1112	0.1134	0.8301	0.8996	29	27
18	Robot2	0.2734	0.2553	0.6740	0.7225	31	28
19	Hillvalley	0.2417	0.2198	0.7141	0.7571	43	39
20	Clean1	0.0086	0.0077	0.9330	0.9984	54	50
21	DNA	0.1301	0.1031	0.8157	0.8877	77	75
22	Semeion	0.088	0.072	0.8871	0.9371	137	135
23	USPS	0.0441	0.0407	0.9004	0.9601	88	86.7
24	Arrhythmia	0.3170	0.3011	0.6347	0.6987	89	86
25	LSVT	0.1242	0.1005	0.8501	0.8997	109	105
26	Madelon	0.258	0.215	0.7411	0.7907	165	166
27	CNAE	0.1274	0.1255	0.8354	0.8811	484	480
28	Colon cancer	0.143	0.100	0.8217	0.8808	635	630
29	Leukemia	0.0249	0.0236	0.9007	0.9763	2065	2053.7
30	Arcene	0.0795	0.0750	0.8365	0.899	3195	3160.85
Average		0.1296	0.1177	0.8244	0.8802	247.3	244.18

by an Intel Core i7-3630QM processor running at 3.2 GHz, and equipped with 8 GB RAM.

A. DATASETS

We utilized a set of 30 datasets obtained from the UCI data repository to assess and verify the effectiveness of BinCOA in comparison to state-of-the-art algorithms. The selection of these datasets was driven by their diverse range of instances and features, providing a thorough evaluation of BinCOA across various challenges. Table 1 offers a concise overview of the examined datasets, encompassing varied class counts, instance quantities, and attribute variations.

B. CONFIGURATION BinCOA PARAMETER

The performance of BinCOA is compared to several other state-of-the-art feature selection methods. Each algorithm undergoes 20 runs, with a maximum iteration limit of 30 and 10 search agents. The chosen classifiers for this study are kNN and SVM. When datasets comprise more than two classes, the 5-NN classifier takes precedence for generating the optimal subset. Thorough trials and runs on diverse datasets are conducted to determine the optimal K value for

kNN. K-fold cross-validation is set at 10 for both kNN and SVM to mitigate overfitting. The parameters for BinCOA are outlined in Table 2.

C. EXPERIMENTAL RESULTS

The experimental process comprises two phases. The initial phase entails a comparison between the proposed BinCOA and the original COA. A comparative analysis is carried out in the subsequent phase between the proposed BinCOA and the latest feature selection algorithms. The experiments in this study are grounded in four primary evaluation measures, as follows:

- Classification accuracy: Classification accuracy refers to the classifier’s precision in determining the most advantageous subset of features
- Average Fitness value: The Average Fitness at each run *n* can be computed as follows:

$$AverageFitness_n = \frac{1}{Maximum\ iteration} \sum_{i=1}^{Maximum\ iteration} Fitness_i \quad (23)$$

where *Fitness_i* is the Fitness at iteration *i*.

TABLE 4. The classification accuracy comparison results with other recent algorithms overall dataset.

NO	Dataset	GWO [49]	EO[50]	MFO[51]	PSO[52]	SSA[53]	HOA[54]	WOA[55]	BinCOA
1	Breast Cancer	0.9681	0.9710	0.9676	0.9694	0.9685	0.9728	0.969	0.9819
2	Fri_c0_1000_10	0.8188	0.8500	0.8221	0.845	0.8275	0.8914	0.8315	0.8997
3	Fri_c0_1000_10	0.8712	0.8910	0.869	0.889	0.8815	0.9053	0.8742	0.9314
4	Vowel	0.9149	0.9426	0.9281	0.9327	0.9295	0.937	0.928	0.9441
5	HeartEW	0.6636	0.6603	0.6591	0.6631	0.6642	0.6731	0.6608	0.6873
6	Congress	0.9588	0.9606	0.9609	0.961	0.9585	0.9676	0.6596	0.9805
7	Hepatitis	0.8509	0.8541	0.4829	0.8506	0.8506	0.8716	0.8429	0.8911
8	Parkinsons	0.8912	0.8910	0.8879	0.8928	0.8933	0.9305	0.8923	0.9484
9	IonosphereEW	0.8843	0.8914	0.8789	0.8884	0.8868	0.9518	0.883	0.9607
10	Dermatology	0.955	0.9748	0.9578	0.9657	0.9588	0.9786	0.9632	0.9997
11	WaveForm	0.789	0.8217	0.7964	0.8042	0.7958	0.853	0.7993	0.8771
12	Lung Cancer	0.6359	0.6375	0.639	0.6546	0.6437	0.6562	0.6281	0.6608
13	Spambase	0.8844	0.9106	0.8913	0.9022	0.889	0.9203	0.8923	0.9414
14	SonarEW	0.8382	0.8403	0.8413	0.8408	0.8326	0.8771	0.8384	0.8833
15	Splice	0.6994	0.7101	0.6974	0.7099	0.7007	0.8066	0.6979	0.8378
16	Movementlibras	0.7705	0.7766	0.7748	0.7794	0.7772	0.783	0.7761	0.7934
17	Robot1	0.8522	0.8625	0.8613	0.8642	0.8608	0.8829	0.8562	0.8996
18	Robot2	0.6436	0.6457	0.6404	0.6446	0.6457	0.7042	0.6457	0.7225
19	Hillvalley	0.5462	0.5490	0.5483	0.5486	0.5551	0.7355	0.5544	0.7571
20	Clean1	0.8815	0.8946	0.8891	0.8884	0.8806	0.9933	0.8862	0.9984
21	DNA	0.8089	0.8274	0.8069	0.8304	0.8126	0.8839	0.8129	0.8877
22	Semeion	0.8913	0.9169	0.9008	0.9	0.9005	0.9254	0.9031	0.9371
23	USPS	0.9576	0.9621	0.9578	0.9589	0.9576	0.9599	0.9587	0.9601
24	Arrhythmia	0.6429	0.6424	0.6404	0.6415	0.6366	0.6829	0.6406	0.6987
25	LSVT	0.6646	0.6619	0.669	0.644	0.6591	0.8801	0.6646	0.8997
26	Madelon	0.7366	0.7525	0.7338	0.7505	0.7437	0.7751	0.74	0.7907
27	CNAE	0.7284	0.8599	0.7612	0.7719	0.7621	0.8739	0.7744	0.8811
28	Colon cancer	0.8064	0.8080	0.8048	0.8048	0.7959	0.879	0.7959	0.8808
29	Leukemia	0.9451	0.9611	0.952	0.9631	0.9555	0.9763	0.9493	0.9763
30	Arcene	0.8425	0.8465	0.8387	0.8452	0.842	0.895	0.8462	0.899
Average		0.8114	0.8258	0.8019	0.8201	0.8155	0.8674	0.8054	0.8802

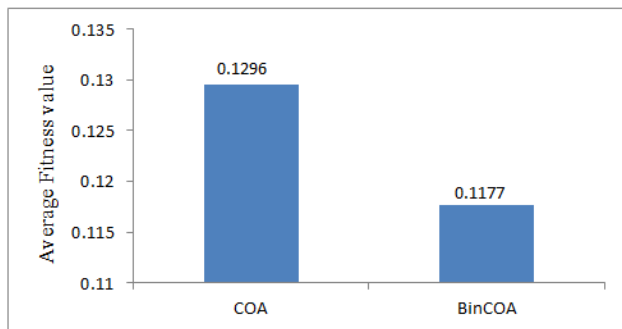


FIGURE 6. The average Fitness value of BinCOA compared to COA over all datasets.

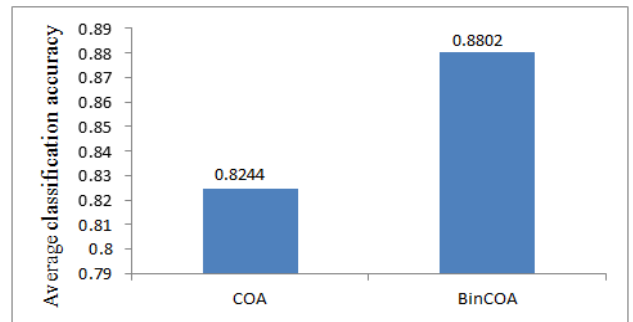


FIGURE 7. Average classification accuracy value of BinCOA compared to COA over all datasets.

- No. of selected feature: denotes the minimum number of features obtained in the optimum solution.

The present experiments in this section examine the impact of combining the refracted opposition-based learning strategy and the crisscross strategy into the performance of the COA algorithm. Table 3 presents comparative studies between the proposed BinCOA and the original COA regarding average fitness, classification accuracy, and No. of selected features. Concerning average fitness, the table

shows that BinCOA consistently outperforms the original COA across all 30 datasets. Table 3 also shows that the BinCOA consistently performs better than the original COA for all 30 datasets in Accuracy classification. The number of selected features for each algorithm is also reported in Table 3. BinCOA achieves the highest ranking in 28 instances out of 30 datasets. Figures 6, 7, and 8 present a comparative analysis between COA and BinCOA, showcasing the overall average fitness value, classification accuracy, and number of selected features across all datasets.

TABLE 5. Average Fitness comparison results with other recent algorithms overall dataset.

NO	Dataset	GWO [49]	EO[50]	MFO[51]	PSO[52]	SSA[53]	HOA[54]	WOA[55]	BinCOA
1	Breast Cancer	0.0352	0.0323	0.0347	0.0335	0.0349	0.0305	0.0348	0.0300
2	Fri_c0_1000_10	0.1838	0.1494	0.1795	0.1563	0.1783	0.1098	0.1714	0.1033
3	Fri_c0_1000_10	0.1314	0.1131	0.1335	0.1127	0.1206	0.0938	0.1262	0.0915
4	Vowel	0.0842	0.0580	0.0769	0.0691	0.0739	0.0641	0.0732	0.0590
5	HeartEW	0.3304	0.3328	0.3311	0.3298	0.3291	0.3167	0.3284	0.3101
6	Congress	0.0426	0.0410	0.0411	0.0397	0.0419	0.0339	0.0416	0.0317
7	Hepatitis	0.1421	0.1404	0.1415	0.14	0.1424	0.121	0.1443	0.117
8	Parkinsons	0.1013	0.0982	0.1029	0.0985	0.1007	0.0629	0.101	0.0610
9	IonosphereEW	0.113	0.1061	0.116	0.1105	0.1103	0.0456	0.1138	0.0434
10	Dermatology	0.0463	0.0294	0.0438	0.0367	0.0439	0.0237	0.0426	0.0219
11	WaveForm	0.2124	0.1843	0.2062	0.1983	0.206	0.1502	0.2038	0.1311
12	Lung Cancer	0.3127	0.2908	0.3053	0.2835	0.3001	0.2737	0.3002	0.2443
13	Spambase	0.1188	0.0942	0.1115	0.1012	0.1137	0.0838	0.1115	0.0668
14	SonarEW	0.1476	0.1458	0.1471	0.1464	0.1509	0.1113	0.1477	0.1003
15	Splice	0.2988	0.2882	0.3024	0.2888	0.2972	0.1795	0.304	0.1665
16	Movementlibras	0.2113	0.2097	0.2147	0.2071	0.2092	0.2059	0.212	0.2022
17	Robot1	0.1346	0.1260	0.1317	0.1239	0.131	0.1169	0.1291	0.1134
18	Robot2	0.3444	0.3407	0.3445	0.3402	0.3441	0.2739	0.3419	0.2553
19	Hillvalley	0.4386	0.4367	0.4396	0.4357	0.4359	0.2299	0.438	0.2198
20	Clean1	0.1116	0.1011	0.1067	0.1041	0.1105	0.0075	0.1084	0.0077
21	DNA	0.1928	0.1737	0.1944	0.1726	0.1867	0.1054	0.1861	0.1031
22	Semeion	0.1063	0.0862	0.1006	0.0988	0.1012	0.078	0.0988	0.072
23	USPS	0.0435	0.0412	0.0433	0.0419	0.0433	0.0416	0.043	0.0407
24	Arrhythmia	0.353	0.3502	0.3544	0.3489	0.3526	0.3056	0.3515	0.3011
25	LSVT	0.3065	0.3069	0.3033	0.3065	0.3061	0.1067	0.3029	0.1005
26	Madelon	0.2625	0.2468	0.2673	0.2486	0.257	0.226	0.2614	0.215
27	CNAE	0.2684	0.1440	0.2383	0.2259	0.2381	0.1292	0.2283	0.1255
28	Colon cancer	0.1518	0.1510	0.1569	0.1546	0.1519	0.103	0.1564	0.100
29	Leukemia	0.0343	0.0318	0.0386	0.0298	0.033	0.0266	0.0358	0.0236
30	Arcene	0.1328	0.1306	0.1315	0.1324	0.1316	0.0785	0.1342	0.0750
Average		0.1797	0.1660	0.1779	0.1705	0.1758	0.1245	0.1757	0.1177

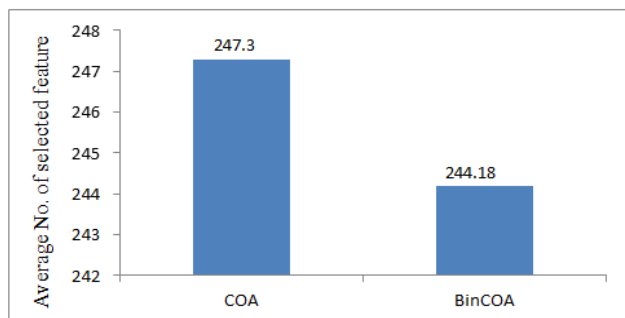


FIGURE 8. Average No. of selected feature value of BinCOA compared to COA over all datasets.

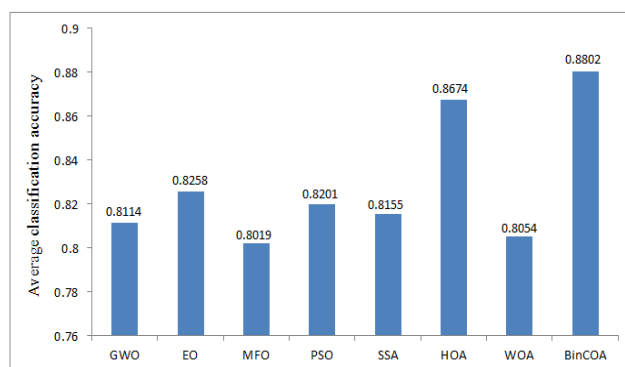


FIGURE 9. Comparison between BinCOA and recent feature selection algorithms in Average Classification accuracy overall datasets.

In order to investigate the performance of the proposed BinCOA algorithm, a comparison of results with the latest feature selection algorithms was conducted. In the comparative results, we use 7 well-known feature selection algorithms: GWO [46], EO [47], MFO [48], PSO [49], SSA [50], HOA [22], and WOA [51]. Tables 4-6 present the numerical outcomes achieved by the proposed BinCOA algorithm in comparison with the latest feature selection algorithms. Table 4 discusses the accuracy of the participants' methodologies and the proposed BinCOA algorithm over 30 iterations on each of the 30 datasets.

Based on the data presented in the table, it can be observed that BinCOA has achieved the highest accuracy

value in 96.6% of the instances, namely in 29 out of the total 30 datasets. Subsequently, it has the highest overall average accuracy across all datasets. Figure 9 depicts a bar chart that illustrates a comparison based on the overall average accuracy. According to the presented data, the figure illustrates that the proposed algorithm exhibits the highest ranking in terms of total average accuracy at 96.6%.

The Fitness value of each of BinCOA and the other algorithms for 30 datasets are listed in table 5. According to the results in table 5, the proposed BinCOA algorithm outperforms other algorithms in 28 out of the total 30 datasets.

TABLE 6. No. of selected feature comparison results with other recent algorithms over all datasets.

NO	Dataset	GWO [49]	EO[50]	MFO[51]	PSO[52]	SSA[53]	HOA[54]	WOA[55]	BinCOA
1	Breast Cancer	7	5	7	5	5	5	5.8	4
2	Fri_c0_1000_10	11	7	10.4	8	9	6.05	9.40	5.5
3	Fri_c0_1000_10	7.5	4.3	7	5	6	3.95	6	3
4	Vowel	12	8.7	11	10	9	8.2	10	7.04
5	HeartEW	10.9	6.9	10.54	8	8.08	6.65	9	5
6	Congress	16	7.06	14.45	10	12.54	6.9	13.08	6
7	Hepatitis	13.51	7.63	12	11	11.87	7.35	12.6	7.35
8	Parkinsons	12.35	7.35	11	11	11	7.35	11.5	7.35
9	IonosphereEW	18	15	17.87	15	16	13.9	16.71	11
10	Dermatology	19	17	19.5	18	19	15.9	18	13
11	WaveForm	31	24	31.5	26	28	22.9	29	21
12	Lung Cancer	27	18	25.63	23	24	16.95	24.06	15
13	Spambase	40.65	31	39.65	35	37	30.3	38.65	28
14	SonarEW	33.54	22.5	30.47	28	28.54	22.5	29.9	20.06
15	Splice	31	23	29.56	27	27.9	22.1	28.59	20
16	Movementlibras	44	35	43	38	40	33	42	30
17	Robot1	39	31.04	36	34	35	29.65	35.47	27
18	Robot2	37	30	33	30	31.54	29.6	32	28
19	Hillvalley	49	43.65	48	45	46.04	41.15	46	39
20	Clean1	65	55	63	58	59	52.25	60	50
21	DNA	95	85	92	90	92	82.85	90.7	75
22	Semeion	165	155	163	159	159.87	152.35	161.08	135
23	USPS	115	105	109	107	108	102.65	109.65	86.7
24	Arrhythmia	106	93.5	103	96	99	91.9	101	86
25	LSVT	136	120	133	124	127	117.15	129.54	105
26	Madelon	210	190	207	197	199	188.5	202	166
27	CNAE	545	510	536	519	526	505.05	530	480
28	Colon cancer	722	656	713.08	660	671.45	644.1	698.45	630
29	Leukemia	2867.47	2687.03	2850.47	2757.04	2780.47	2369.9	2830.47	2053.7
30	Arcene	3680	3630.75	3664.65	3698.37	3770	3530.75	3815.14	3160.85
	Average	305.53	287.71	302.39	295.08	299.91	272.22	304.85	244.18

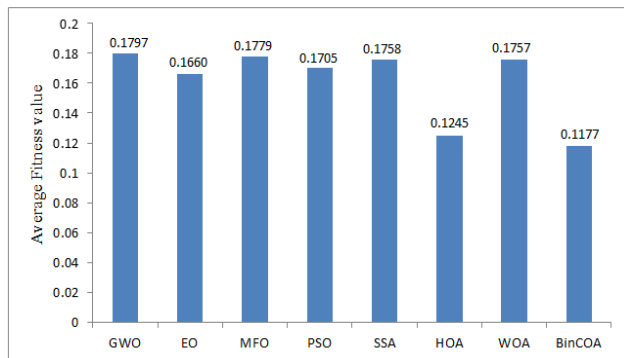


FIGURE 10. Comparison between BinCOA and recent feature selection algorithms in Average Fitness over all datasets.

The bar chart in Figure 10 compares the average fitness value for BinCOA and the other algorithms. The IBEVO algorithm has superior performance in achieving the minimal average fitness value (0.1177) across all datasets, and the HOA algorithm comes in second with a value of (0.1245) as shown in figure 10.

In addition to maximizing classification accuracy, minimizing the number of selected features is also considered desirable.

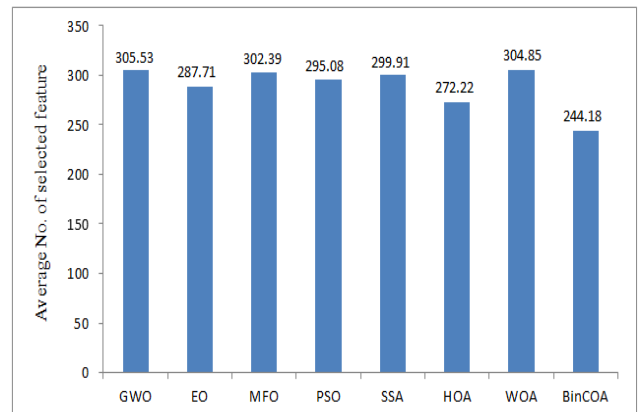


FIGURE 11. Comparison between BinCOA and recent feature selection algorithms in Average Fitness over all datasets.

The number of selected features for all datasets is reported in table 6. The proposed BinCOA achieves the minimum number of selected features for 30 of the 30 total datasets. The IBEVO demonstrates strong size reduction capabilities, obtaining the smallest average selection size (244.18) across all datasets. The HOA algorithm is ranked second with a value of (272.22), as depicted in Figure 11. The Wilcoxon signed

TABLE 7. The Wilcoxon rank sum test results.

NO	Dataset	BinCOA VS GWO	BinCOA VS EO	BinCOA VS MFO	BinCOA VS PSO	BinCOA VS SSA	BinCOA VS HOA	BinCOA VS WOA
1	Breast Cancer	5.874e-38	8.5467e-41	3.7504e-33	7.0441e-33	6.5652e-26	5.6804e-23	6.741e-44
2	Fri_c0_1000_10	3.4151e-41	1.4787e-33	3.1052e-37	2.3351e-41	6.7034e-33	3.4707e-16	2.0671e-41
3	Fri_c0_1000_10	7.7481e-33	6.3748e-36	8.8341e-19	6.7447e-19	1.7771e-17	3.7747e-09	6.7447e-19
4	Vowel	1.8346e-38	6.3004e-09	2.7047e-38	3.1007e-43	3.65447e-19	5.0841e-13	3.1007e-41
5	HeartEW	5.1477e-35	7.3101e-07	7.551e-40	1.8559e-07	8.7440e-13	3.774e-06	8.8659e-07
6	Congress	1.4178e-45	9.7343e-26	3.8730e-43	1.5882e-40	4.5693e-41	7.5416e-38	6.5502e-43
7	Hepatitis	2.3398e-37	7.2607e-38	8.740e-21	3.7484e-33	3.74861e-47	7.741e-33	4.7484e-34
8	Parkinsons	6.7447e-08	1.4903e-43	4.7441e-11	5.2447e-41	1.8154e-36	7.1370e-41	5.2687e-40
9	IonosphereEW	3.7411e-35	8.9551e-27	8.7474e-43	6.441e-38	1.5588e-35	3.1183e-33	6.7441e-37
10	Dermatology	4.5876e-28	3.6284e-18	3.7701e-08	1.2830e-21	3.3602e-40	1.8934e-18	1.2830e-28
11	WaveForm	1.5477e-40	57444e-39	7.6746e-34	6.8018e-41	3.4141e-45	9.9604e-41	6.8078e-44
12	Lung Cancer	8.4411e-35	5.5581e-33	7.7444e-14	4.2803e-45	3.6874e-33	8.4335e-11	3.44801e-43
13	Spambase	3.2241e-38	6.7468e-27	5.3412e-23	1.5505e-48	1.78744e-13	2.772e-23	1.3551e-34
14	SonarEW	3.1047e-44	6.678e-38	3.7141e-40	3.0478e-30	4.8501e-47	7.6841e-09	3.1047e-40
15	Splice	4.4568e-41	4.8443e-43	3.7481e-35	7.5474e-38	3.877e-47	8.5844e-33	4.4018e-41
16	Movementlibras	3.9634e-39	6.4704e-19	8.5543e-39	5.4225e-10	5.4405e-44	3.8542e-41	3.0053e-34
17	Robot1	1.4716e-37	7.7477e-41	3.5823e-23	2.5446e-44	6.6357e-38	1.8837e-37	1.85476e-35
18	Robot2	2.7418e-43	7.7778e-43	7.9347e-43	3.3057e-31	6.3001e-38	3.3108e-32	1.7447e-44
19	Hillvalley	1.4718e-25	5.7448e-25	8.3797e-30	3.4406e-41	3.7742e-37	4.3441e-27	6.1453e-07
20	Clean1	7.1456e-37	2.6118e-32	7.3874e-39	3.2474e-31	6.043e-14	3.5747e-29	6.8457e-41
21	DNA	3.5547e-08	4.0147e-32	7.4742e-33	3.3636e-37	7.7741e-43	1.3447e-41	9.4787e-33
22	Semeion	4.5187e-41	9.5411e-39	2.7410e-11	1.7368e-41	4.6314e-44	7.7784e-44	1.3985e-43
23	USPS	1.5474e-16	8.1418e-45	2.5477e-35	3.7400e-41	7.7744e-17	6.7775e-40	3.3344e-09
24	Arrhythmia	2.4708e-36	8.4018e-41	7.113e-27	1.4114e-44	3.7473e-36	6.5741e-17	2.3401e-08
25	LSVT	4.1305e-33	7.4041e-15	9.7404e-40	9.6507e-38	7.844e-33	3.3444e-44	3.7013e-23
26	Madelon	7.1763e-36	3.78741e-10	2.178e-33	2.6714e-37	1.4131e-33	3.8354e-20	4.6714e-37
27	CNAE	6.51074e-19	3.7418e-40	5.4774e-37	3.7400e-38	6.7874e-18	6.1875e-43	3.1440e-44
28	Colon cancer	2.4787e-36	1.7418e-41	3.5563e-30	7.3014e-43	2.8003e-38	4.7084e-15	1.3544e-36
29	Leukemia	9.1345e-41	3.0171e-19	1.4704e-41	3.1407e-36	7.5518e-39	3.8754e-44	9.1547e-38
30	Arcene	4.1743e-31	7.8311e-11	2.0727e-36	5.6714e-31	4.7744e-31	1.8474e-27	3.6714e-36

rank-sum test is a statistical technique employed to evaluate the similarity or dissimilarity between two groups. This test analyzes the differences within pairs of groups to determine if they are statistically significantly distinct. In our analysis, the Wilcoxon rank-sum test, conducted at a 5% significance level, compares the results of the BinCOA algorithm to six prominent recent feature selection metaheuristic algorithms across the 30 standard datasets. Table 7 displays the p-values obtained from this test. Upon examination of the data in the table, it becomes evident that all p-values for the compared algorithms fall below the 5% significance level. This result provides compelling evidence to reject the null hypothesis. Consequently, it can be inferred that the binary BinCOA method surpasses all other comparative algorithms.

VI. CONCLUSION

A novel Crayfish Optimization Algorithm (BinCOA) provides for feature selection problems in the present study. The original COA is enhanced by incorporating both the refracted opposition-based learning strategy and the crisscross strategy, leading to improved performance. The k-nearest neighbors (kNN) or support vector machine (SVM) classifier has

been found to produce high-quality solutions when used in conjunction with the BinCOA algorithm. Furthermore, these classifiers have proven their ability to learn effectively from the provided training data. The application of k-fold cross-validation is a highly effective approach for addressing the concern of overfitting. In order to promote traversal and variety, the population is initialized using the refracted opposition-based learning technique. It has been discovered that applying the crisscross technique improves optimization accuracy to some extent. It also facilitates a more thorough investigation of possible answers and enhances the algorithm’s utilization of the search space. Thirty datasets are used to evaluate the proposed algorithm, and the results are compared with seven well-known feature selection algorithms. The contrasting experiments and the mentioned results demonstrate the superiority of BinCOA over recent feature selection algorithms. Furthermore, the significance of the proposed algorithm is assessed by the utilization of the Wilcoxon rank-sum test. The statistical findings indicate that the proposed algorithm demonstrates superior performance when compared to the most recent feature selection algorithms.

REFERENCES

- [1] L.-P. Jing, H.-K. Huang, and H.-B. Shi, "Improved feature selection approach TFIDF in text mining," in *Proc. Int. Conf. Mach. Learn. Cybern.*, vol. 2, Nov. 2002, pp. 944–946.
- [2] P. Wongthongtham, J. Kaur, V. Potdar, and A. Das, "Big data challenges for the Internet of Things (IoT) paradigm," in *Connected Environments for the Internet of Things: Challenges and Solutions*, 2017, pp. 41–62.
- [3] D. A. Elmanakhly, M. M. Saleh, and E. A. Rashed, "An improved equilibrium optimizer algorithm for features selection: Methods and analysis," *IEEE Access*, vol. 9, pp. 120309–120327, 2021.
- [4] K. Huang and S. Aviyente, "Wavelet feature selection for image classification," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1709–1720, Sep. 2008.
- [5] B. Ghaddar and J. Naoum-Sawaya, "High dimensional data classification and feature selection using support vector machines," *Eur. J. Oper. Res.*, vol. 265, no. 3, pp. 993–1004, Mar. 2018.
- [6] H. Faris, M. M. Mafarja, A. A. Heidari, I. Aljarah, A. M. Al-Zoubi, S. Mirjalili, and H. Fujita, "An efficient binary salp swarm algorithm with crossover scheme for feature selection problems," *Knowl.-Based Syst.*, vol. 154, pp. 43–67, Aug. 2018.
- [7] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4–37, Jan. 2000.
- [8] B. V. Dasarathy, "Nearest neighbor (NN) norms: NN pattern classification techniques," *IEEE Comput. Soc. Tutorial*, 1991.
- [9] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary ant lion approaches for feature selection," *Neurocomputing*, vol. 213, pp. 54–65, Nov. 2016.
- [10] C. Kuzudisli, B. Bakir-Gungor, N. Bulut, B. Qaqish, and M. Yousef, "Review of feature selection approaches based on grouping of features," *PeerJ*, vol. 11, Jul. 2023, Art. no. e15666.
- [11] A. M. Khalid, H. M. Hamza, S. Mirjalili, and K. M. Hosny, "BCOVIDOA: A novel binary coronavirus disease optimization algorithm for feature selection," *Knowl.-Based Syst.*, vol. 248, Jul. 2022, Art. no. 108789.
- [12] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," *Neurocomputing*, vol. 172, pp. 371–381, Jan. 2016.
- [13] H. Jia, H. Rao, C. Wen, and S. Mirjalili, "Crayfish optimization algorithm," *Artif. Intell. Rev.*, vol. 56, no. 2, pp. 1919–1979, Nov. 2023.
- [14] R. V. Rao, V. J. Savsani, and D. P. Vakharia, "Teaching-learning-based optimization: A novel method for constrained mechanical design optimization problems," *Comput.-Aided Design*, vol. 43, no. 3, pp. 303–315, Mar. 2011.
- [15] A. Kaveh and N. Farhoudi, "A new optimization method: Dolphin echolocation," *Adv. Eng. Softw.*, vol. 59, pp. 53–70, May 2013.
- [16] N. Hansen, S. D. Müller, and P. Koumoutsakos, "Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES)," *Evol. Comput.*, vol. 11, no. 1, pp. 1–18, Mar. 2003.
- [17] E. Rashedi, H. Nezamabadi-Pour, and S. Saryzadi, "GSA: A gravitational search algorithm," *Inf. Sci.*, vol. 179, no. 13, pp. 2232–2248, Jun. 2009.
- [18] P. Agrawal, T. Ganesh, and A. W. Mohamed, "A novel binary gaining-sharing knowledge-based optimization algorithm for feature selection," *Neural Comput. Appl.*, vol. 33, no. 11, pp. 5989–6008, 2021.
- [19] S. Hosseini and A. Al Khaled, "A survey on the imperialist competitive algorithm Metaheuristic: Implementation in engineering domain and directions for future research," *Appl. Soft Comput.*, vol. 24, pp. 1078–1094, Nov. 2014.
- [20] H. C. Kuo and C. H. Lin, "Cultural evolution algorithm for global optimizations and its applications," *J. Appl. Res. Technol.*, vol. 11, no. 4, pp. 510–522, Aug. 2013.
- [21] R. Moghdani and K. Salimifard, "Volleyball premier league algorithm," *Appl. Soft Comput.*, vol. 64, pp. 161–185, Mar. 2018.
- [22] D. A. Elmanakhly, M. Saleh, E. A. Rashed, and M. Abdel-Basset, "BinHOA: Efficient binary horse herd optimization method for feature selection: Analysis and validations," *IEEE Access*, vol. 10, pp. 26795–26816, 2022.
- [23] D. Rodrigues, L. A. M. Pereira, T. N. S. Almeida, J. P. Papa, A. N. Souza, C. C. O. Ramos, and X.-S. Yang, "BCS: A binary cuckoo search algorithm for feature selection," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2013, pp. 465–468.
- [24] M. M. Mafarja, D. Eleyan, I. Jaber, A. Hammouri, and S. Mirjalili, "Binary dragonfly algorithm for feature selection," in *Proc. Int. Conf. New Trends Comput. Sci. (ICTCS)*, Oct. 2017, pp. 12–17.
- [25] D. Rodrigues, X. S. Yang, A. N. De Souza, and J. P. Papa, "Binary flower pollination algorithm and its application to feature selection," *Recent Adv. Swarm Intell. Evol. Comput.*, vol. 585, pp. 85–100, Dec. 2014.
- [26] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: Novel initialisation and updating mechanisms," *Appl. Soft Comput.*, vol. 18, pp. 261–276, May 2014.
- [27] Q. Al-Tashi, S. J. A. Kadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Binary optimization using hybrid grey wolf optimization for feature selection," *IEEE Access*, vol. 7, pp. 39496–39508, 2019.
- [28] M. Kumar, M. Husain, N. Upreti, and D. Gupta, "Genetic algorithm: Review and application," *SSRN Electron. J.*, 2010.
- [29] H. Chen, W. Jiang, C. Li, and R. Li, "A heuristic feature selection approach for text categorization by using chaos optimization and genetic algorithm," *Math. Problems Eng.*, vol. 2013, pp. 1–6, May 2013.
- [30] Y. Zhang, D.-W. Gong, X.-Z. Gao, T. Tian, and X.-Y. Sun, "Binary differential evolution with self-learning for multi-objective feature selection," *Inf. Sci.*, vol. 507, pp. 67–85, Jan. 2020.
- [31] D. Simon, "Biogeography-based optimization," *IEEE Trans. Evol. Comput.*, vol. 12, no. 6, pp. 702–713, Mar. 2008.
- [32] S. Khalilpourazari, B. Naderi, and S. Khalilpourazary, "Multi-objective stochastic fractal search: A powerful algorithm for solving complex multi-objective optimization problems," *Soft Comput.*, vol. 24, no. 4, pp. 3037–3066, Feb. 2020.
- [33] H. Shareef, A. A. Ibrahim, and A. H. Mutlag, "Lightning search algorithm," *Appl. Soft Comput.*, vol. 36, pp. 315–333, Nov. 2015.
- [34] S. Mirjalili, S. M. Mirjalili, and A. Hatamlou, "Multi-verse optimizer: A nature-inspired algorithm for global optimization," *Neural Comput. Appl.*, vol. 27, no. 2, pp. 495–513, Feb. 2016.
- [35] H. Abedinpourshorban, S. M. Shamsuddin, Z. Beheshti, and D. N. A. Jawawi, "Electromagnetic field optimization: A physics-inspired metaheuristic optimization algorithm," *Swarm Evol. Comput.*, vol. 26, pp. 8–22, Feb. 2016.
- [36] F. A. Hashim, E. H. Houssein, M. S. Mabrouk, W. Al-Atabany, and S. Mirjalili, "Henry gas solubility optimization: A novel physics-based algorithm," *Future Gener. Comput. Syst.*, vol. 101, pp. 646–667, Dec. 2019.
- [37] M. Taradeh, M. Mafarja, A. A. Heidari, H. Faris, I. Aljarah, S. Mirjalili, and H. Fujita, "An evolutionary gravitational search-based feature selection," *Inf. Sci.*, vol. 497, pp. 219–239, Sep. 2019.
- [38] F. S. Hosseini, B. Choubin, A. Mosavi, N. Nabipour, S. Shamshirband, H. Darabi, and A. T. Haghghi, "Flash-flood hazard assessment using ensembles and Bayesian-based machine learning models: Application of the simulated annealing feature selection method," *Sci. Total Environ.*, vol. 711, Apr. 2020, Art. no. 135161.
- [39] A. Faramarzi, M. Heidarinejad, B. Stephens, and S. Mirjalili, "Equilibrium optimizer: A novel optimization algorithm," *Knowl.-Based Syst.*, vol. 191, Mar. 2020, Art. no. 105190.
- [40] S. Ahmed, K. K. Ghosh, S. Mirjalili, and R. Sarkar, "AIEOU: Automata-based improved equilibrium optimizer with U-shaped transfer function for feature selection," *Knowl.-Based Syst.*, vol. 228, Sep. 2021, Art. no. 107283.
- [41] E. R. Larson and J. D. Olden, "The state of crayfish in the Pacific Northwest," *Fisheries*, vol. 36, no. 2, pp. 60–73, Feb. 2011.
- [42] F. Zhao, L. Zhang, Y. Zhang, W. Ma, C. Zhang, and H. Song, "An improved water wave optimisation algorithm enhanced by CMA-ES and opposition-based learning," *Connection Sci.*, vol. 32, no. 2, pp. 132–161, Apr. 2020.
- [43] M. M. Mafarja and S. Mirjalili, "Hybrid whale optimization algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302–312, Oct. 2017.
- [44] A.-B. Meng, Y.-C. Chen, H. Yin, and S.-Z. Chen, "Crisscross optimization algorithm and its application," *Knowl.-Based Syst.*, vol. 67, pp. 218–229, Sep. 2014.
- [45] F. Pernkopf, "Bayesian network classifiers versus selective k -NN classifier," *Pattern Recognit.*, vol. 38, no. 1, pp. 1–10, Jan. 2005.

- [46] E.-S.-M. El-Kenawy, M. M. Eid, M. Saber, and A. Ibrahim, "MbGWO-SFS: Modified binary Grey Wolf optimizer based on stochastic fractal search for feature selection," *IEEE Access*, vol. 8, pp. 107635–107649, 2020.
- [47] Y. Gao, Y. Zhou, and Q. Luo, "An efficient binary equilibrium optimizer algorithm for feature selection," *IEEE Access*, vol. 8, pp. 140936–140963, 2020.
- [48] R. Khurma, I. Aljarah, and A. Shariéh, "An efficient moth flame optimization algorithm using chaotic maps for feature selection in the medical applications," in *Proc. 9th Int. Conf. Pattern Recognit. Appl. Methods*, 2020, pp. 175–182.
- [49] S. B. Sakri, N. B. A. Rashid, and Z. M. Zain, "Particle swarm optimization feature selection for breast cancer recurrence prediction," *IEEE Access*, vol. 6, pp. 29637–29647, 2018.
- [50] S. S. Shekhawat, H. Sharma, S. Kumar, A. Nayyar, and B. Qureshi, "BSSA: Binary salp swarm algorithm with hybrid data transformation for feature selection," *IEEE Access*, vol. 9, pp. 14867–14882, 2021.
- [51] M. Tubishat, M. A. M. Abushariah, N. Idris, and I. Aljarah, "Improved whale optimization algorithm for feature selection in Arabic sentiment analysis," *Appl. Intell.*, vol. 49, no. 5, pp. 1688–1707, May 2019.

AHMED SALEM AL-ERAQI is currently an Assistant Professor with the Department of Computer Science and Engineering, Faculty of Engineering, Aden University. His research interests include computational intelligence, machine learning, social analytics, IT, enterprise systems, and mobile HCI.B.

ISLAM S. FATHI received the B.Sc. and M.Sc. degrees in mathematics and computer sciences from the Faculty of Science, Zagazig University, Egypt, in 2013 and 2019, respectively, and the Ph.D. degree in computer science from the Faculty of Science, Suez Canal University, Egypt, in 2023. His research interests include artificial intelligence, signal processing, metaheuristic optimization, machine learning, deep learning, bioinformatics, and the Internet of Things.

• • •



NABILA H. SHIKOUN received the dual B.Sc. degree in computer and automatic control and computer engineering from Ain Shams University, Egypt, in 1981, the M.Sc. degree in computers and systems engineering from Helwan University, in 2006, and the Ph.D. degree in computer and systems engineering from Al-Azhar University, in 2014. Her research interests include climate change, signal processing, metaheuristics, image processing, and bioinformatics.