STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–8.
Published online in International Academic Press (www.IAPress.org)

Feature selection using binary Harris Hawks optimization algorithm to improve K-Means clustering

Shaymaa Haleem Ibrahem [1], Ammar Saad Abduljabbar [2,*], Omar Saber Qasim [2]

[1] *Department of Arts, Mosul University, Iraq*
[2] *Department of Mathematics, Mosul University, Iraq*

Abstract
This study aims to explore the suitability of adopting K-means clustering for the categorization of five disaggregate datasets that undergo feature selection employing a Binary Harris Hawks Optimization Algorithm (BHHOA).
First, a feature selection technique is applied using BHHOA to identify the most important features from each dataset prior to executing dimensionality reduction and enhancing the quality of the collected data. In the next step, the K-means clustering algorithm is applied to the fine-tuned datasets to form a meaningful number of clusters.
The evaluation of K-means clustering considered the effectiveness of the clustering algorithm in terms of accuracy and the selected feature set. Comparing the results with those obtained using the same set of features selected by other methods, it is evident that BHHOA improves feature selection and thereby enhances clustering performance. This confirms its capability to handle large datasets with high dimensionality.
The outcomes show that using the proposed approach—consisting of BHHOA for feature selection followed by K-means clustering—can significantly improve the classification performance of high-dimensional datasets.

Keywords Clustering; K-means; Binary Harris Hawks optimization algorithm; Unsupervised feature selection.

DOI: 10.19139/soic-2310-5070-2653

1. Introduction

Clustering data is a simple activity in scientific and applied areas and allows for classifying large volumes of data of similar objects. To further advance the procedure, this paper recommends implementing an integrated framework which would reduce dimensions as well as clustering. The rationale behind this framework is that subject to enhancing the latent representation, the learning objective is to enhance performance of the K-means algorithm, particularly in cases involving complicated and non-linear data structures. Currently, natural meta-heuristic approaches have become important in high-dimensional feature selection problems. The cooperative hunting style of Harris hawks, which has demonstrated encouraging results due to its broad exploration and exploitation potential, is the basis for one of these algorithms, the Harris Hawks Optimisation (HHO) algorithm. Numerous binary variations of HHO have been created to make the technique appropriate for discrete optimisation problems in medical applications where data sets need to be optimally and effectively segmented [1]. In this context, the binary Harris-Hawks optimisation (BHHO) model was proposed by Dong et al. (2024) for feature selection in cervical cancer diagnosis. Using a rank-based selection approach, the model demonstrated higher stability and accurate classification than the traditional envelope-based methods [2]. Similarly, Alapol et al. (2021) made a complete review of HHO and its chemists, reviewing over 60 works in the medical field, engineering,

*Correspondence to: Ammar Saad Abduljabbar (Email: ammarsaad86@uomosul.edu.iq). Department of Mathematics, Mosul University, Mosul, Nineveh, Iraq (41002).

ISSN 2310-5070 (online) ISSN 2311-004X (print)

and the analysis of images. The paper reviewed findings about the effectiveness of HHO and identified some of the problems thereof, namely, the reliability of HHO in convergence and sensitivity to the parameters of control, which binary or hybrid models like BHHO are intended to address[3]. Also, Issa et al. (2022) introduced a hybrid scheme, based on the combination of HHO and negative swarm optimization (SSO), called HHOSSA, to COVID-19 detection with chest X-ray images in 2022. The accuracy of classification obtained by using their method was over 96 percent, which explored the potential of the HHO-based models in medical image analysis as well as the reduction of features in high dimensions[4]. In addressing complicated continuous optimization issues, Heidari and Mirjalili (2019) have presented the original HHO algorithm based on the simulated dynamic action of the Harris hawks when hunting. Its binary variant, BHHOA, has since shown good competitive capability on large-scale binary optimization problems. Therefore, BHHOA is a good choice for selecting the best feature set before clustering due to its capabilities [5] [6] [7]. The paper introduces a new feature selection method which is a combination of BHHOA algorithm and K-means algorithm used as clustering technique. The algorithm has been used on five standard datasets in which BHHOA algorithm is used to select the best features in the dataset and then Clustering using K-means is done on the best features. The success of such a technique can be measured by clustering accuracy and quality of features, and the effectiveness of this kind of techniques can be measured in terms of enhancing clustering outcomes in high-dimensional data scenarios. The remaining content of the paper can be organised as follows: In the second section K-means clustering technique will be reviewed and the Harris Hawks Optimization algorithm explained, in the third section. Part four is devoted to the introduction of the suggested BHHOA-K-means procedure in the context of features selection, grouping, performance measures, and merits of it. Last, the conclusion is given in the fifth area. The measures of evaluation that can be applied to check the effectiveness of the method include clustering accuracy as well as feature quality that prove that this method can actually enhance the success of clustering with high-dimensional data.

2. K-means

K-means clustering, a widely used unsupervised machine learning technique, segments a dataset into K distinct groups based on feature similarity. Its principles, functionality, and use cases are extensively explored. The method focuses on grouping similar data points while reducing the variance within each cluster. Since it works without labelled data, K-means is ideal for situations where the data's underlying structure is unknown.

The "K" in K-means denotes the number of clusters, which must be determined by the user before the algorithm is applied [8].

The stages taken by the K-means algorithm to accomplish clustering are as follows:

1. **Initialization:** Choose K starting centroids at random from the dataset. These centroids serve as the starting points for each cluster.
2. **Assignment step:** Each data point is assigned to the closest centroid using a distance measure, usually the Euclidean distance. K clusters are successfully formed in this stage.
3. **Update step:** Calculate new centroids by summing all of the data points that are assigned to each cluster. Each centroid's location is revised accordingly.
4. **Convergence check:** Carry out the assignment and update procedures once again until either the centroids stop changing noticeably, signifying that clustering has stabilized, or a certain number of iterations 1,2,3 has been reached [9] [19].

Through examining the visual plots, we can understand the previously mentioned steps:

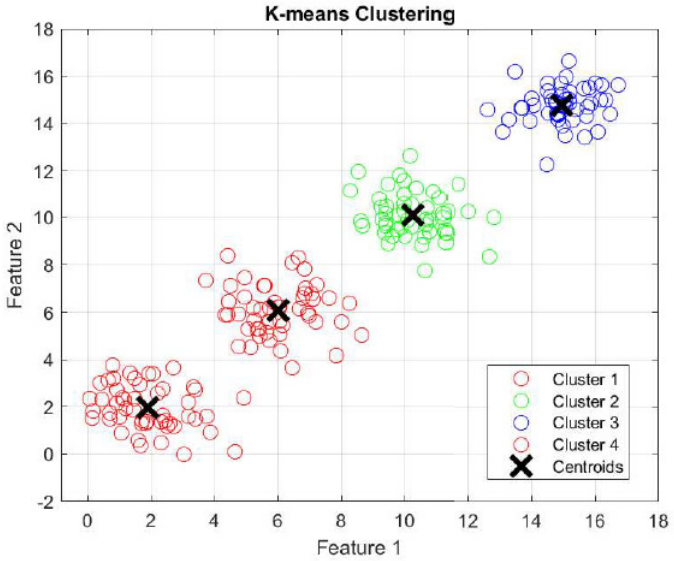

Figure 1. A representation of the K-means clustering

K-means offers many advantages, including ease of implementation and rapid convergence of results, especially when K is small. It is flexible, as you can vary it by using different distance measures and starting strategies. However, it has some drawbacks, including that choosing K in advance can be difficult, and the initial center of gravity position can affect the final clusters. Incorrect choices can lead to suboptimal solutions. Furthermore, K-means assumes an equal-sized spherical cluster, although this may not be true for all data sets. Finally, K-means is a powerful technique for identifying patterns in unlabeled data; however, its limitations and parameter choices must be carefully considered [10][17].

3. Harris Hawks optimization algorithm (HHOA)

The Harris Hawks Optimization Algorithm (HHOA), as proposed by [5]. It is based on artificial mimicking of the processes that occur during hunting and rabbit catching of Harris hawks in nature. The optimization process of HHOA is in three phases, which aim at reaching the best solution to every problem that may arise. These phases include exploration, the transition from exploration to exploitation, and the actual process of exploitation.

3.1. Exploration Phase

The exploration phase is used to describe the condition which is somewhat resembling to that when the Harris hawk is in a position where it is unable to locate the prey properly. When that happens, the hawks cease flying and begin hovering just to survey and point out new targets. In the HHOA, the candidate solutions are the hawks, while the best solution so far at the various steps is the prey. The hawks then perch at a new location just anywhere they wish and wait for prey using two operators that are chosen at random with probability q[2,3] [18].

In mathematical terms this process is represented by

$$x^{(t+1)} = \begin{cases} x^t_{\text{rand}} - r_1 \left| x^t_{\text{rand}} - 2r_2 x^t \right|, & \text{if } q \geq 0.5 \\ (x^t_{\text{prey}} - x^t_m) - r_3(L_b + r_4(U_b - L_b)), & \text{if } q < 0.5 \end{cases} \tag{1}$$

where $x^{(t+1)}$ is the vector representing the hawks' position in the next iteration, x^t_{prey} represents the position of the intended rabbit, and x^t_{rand} is the location of a hawk that is selected at random from the current squad.

r_1, r_2, r_3, and r_4 are random numbers [11]. L_b and U_b are the lower and upper bounds of the search space, respectively.

The current population of hawks' average position is determined by the following equation:

$$x_m^t = \frac{1}{nh} \sum_{i=1}^{nh} x_i^t \tag{2}$$

Where x_i^t denotes the total number of team members and nh represents the position of each hawk on the team [11].

3.2. Transition Phase

The HHO algorithm moves from the exploration phase to the exploitation phase based on the energy level of the prey (escape energy) E. The energy reduction of the prey is defined as

$$E = 2E_0 \left(1 - \frac{t}{t_{\max}}\right) \tag{3}$$

Where $t_{\max}$ indicates the maximum number of iterations, and E_0, which varies inside $(-1, 1)$ each iteration at random. This value indicates if the prey is physically flagged for $-1 \leq E_0 < 0$ or if $0 \leq E_0 < 1$ is strengthened. In the case $|E| \geq 1$, HHOA will then investigate the search space; otherwise, it will enter the exploitation phase [12].

3.3. Exploitation Phase

The selection of the type of besiege to capture the prey is taken $|E|$ into consideration during the exploitation phase. A soft one is taken when $|E| \geq 0.5$, and the hard one is taken when $|E| < 0.5$ [13] [14].

The two strategies of "soft besiege" and "hard besiege" encourage this process. In the soft besiege strategy, $r \geq 0.5$ and $|E| \geq 0.5$ (where r denotes the capacity of the prey to flee). This indicates that the prey still has sufficient energy to escape, so the Harris hawks update the solution by choosing the best answer from the population. The following formula can be used to formulate this:

$$x^{(t+1)} = \Delta x^t - E \cdot |J \cdot x_{\text{prey}}^t - x^t| \tag{4}$$

Where $\Delta x^t = x_{\text{prey}}^t - x^t$, $J = 2(1 - r_5)$, which represents the prey's jump severity during the escape phase, and r_5 is a random number between 0 and 1.

On the other hand, in a hard besiege strategy, $r \geq 0.5$ and $|E| < 0.5$, which indicates that the prey is worn out and lacks the energy to get away. The Harris's hawk's most recent location is described as

$$x^{(t+1)} = x_{\text{prey}}^t - E \cdot |\Delta x^t| \tag{5}$$

In case of $r < 0.5$ and $|E| \geq 0.5$, Harris's hawk gradually chooses the best possible dive to catch the prey competitively, a tactic known as soft besiege with progressively rapid dives [15]. The hawk's new location is then represented mathematically as

$$\gamma = x_{\text{prey}}^t - E \cdot |J \cdot x_{\text{prey}}^t - x^t| \tag{6}$$

The Harris' hawk can dive by

$$Z = \gamma + S \times \text{Levy}(D) \tag{7}$$

Where D is the dimension of the problem, S is a random vector of size $1 \times D$, and Levy is the Levy flight function, which is calculated as

$$\text{Levy}(D) = 0.01 \times \frac{\mu}{|\delta|^{1/\beta}} \left(\frac{\Gamma(1+\beta) \cdot \sin(\pi\beta/2)}{\Gamma\left(\frac{1+\beta}{2}\right) \cdot \beta \cdot 2^{(1+\beta)/2}} \right)^{1/\beta} \tag{8}$$

Where μ and δ are random values from $(0, 1)$ and β is a constant, and its value is 1.5 [11, 16]. In this phase, the position of the Harris' hawk is updated as

$$x^{(t+1)} = \begin{cases} \gamma & \text{if Fitness}(\gamma) < \text{Fitness}(x^t) \\ Z & \text{if Fitness}(Z) < \text{Fitness}(x^t) \end{cases} \tag{9}$$

Where Fitness(x^t) is the fitness function.

3.4. Binary Harris hawks optimizer

The Binary Harris Hawks Optimizer (BHHO) is a new algorithm that is based on the original Harris Hawks Optimizer (HHO). The binary version is largely designed to solve binary optimization problems, which involve decision variables encoded with binary values 0 and 1, corresponding to the possible states or decisions in the problem domain. BHHO conducts ongoing position updates. Normalizing the shape and then using a transfer function to make it binary is one method to accomplish this. Each hawk solution's fitness is assessed according to how well it resolves the optimization issue.

The transfer function $T(X)$ is applied to the continuous position vectors to transform them into binary vectors. The most commonly used transfer function is the sigmoid function. For example:

$$T(X) = \frac{1}{1 + e^{-X}} \tag{10}$$

Where X represents a hawk's location in the continuous space.

The binary position $T(X)$ is then determined using a threshold:

$$T(X) = \begin{cases} 1 & \text{if } T(X) \geq r \\ 0 & \text{if } T(X) < r \end{cases} \tag{11}$$

Where r is a random number that is uniformly distributed in the range $[0, 1]$. The Harris binary optimization algorithm is an exploratory algorithm that gradually moves toward exploitation by increasing the positions of falcons in the search space as prey energy decreases. One advantage of the BHHO algorithm is that it balances exploration and exploitation. The BHHO algorithm keeps exploration of the search space and exploitation of optimal solutions under control, preventing local optimal solutions and enhancing convergence. The algorithm can be easily implemented and modified to address a variety of binary optimization problems, including feature selection, binary classification, and combinatorial problems[7].

4. The proposed algorithm

The proposed algorithm combines binary Harris-Hawks optimization (BHHOA) with k-means clustering to address the challenges of high-dimensional data through a hybrid approach that combines feature selection and clustering optimization. This method improves clustering accuracy and computational efficiency by strategically reducing dimensionality while preserving important data patterns. The proposed methodology operates by including feature selection in stage 1 and clustering in stage 2. The workflow is defined as follows:

1. **High-dimensional data input:** The process begins with high-dimensional raw data that requires segmentation, such as datasets derived from medical sources, images, or text.
2. **Feature selection via binary Harris-Hawks optimization (BHHOA):** BHHOA uses binary optimization to identify essential features, reducing dimensionality while preserving discriminative features. This stage optimizes feature subsets by balancing exploration (exhaustive search) and exploitation (local optimization). Using BHHOA results in a dimensionally reduced dataset, which is optimized by reducing redundancy, effectively reducing the computational complexity of subsequent tasks.

3. **K-means clustering:** K-means starts out by randomly choosing k centroids from the set of k centroids with the reduced dimensions of data. The data is also assigned to the closest centroid using Euclidean distance, and then the Euclidean distances are used to update the centroids by calculating their means. This is repeated until convergence, i.e., when the centroid becomes acceptable and internal variance is minimised.
   This study used a preset number of priorities ($K = 3$) when clustering, according to the conclusions made during the course of preliminary analysis by the Calinski–Harabasz (CH) index. The algorithm was given many instances in order to make it less susceptible to random initialization. The final clustering solution was chosen accordingly.
4. **Performance evaluation:** The effectiveness of the clustering process is evaluated using metrics such as the Silhouette score, and computation time, which collectively assess the quality and efficiency of the clustering. An algorithm exits when a stop condition is satisfied, e.g. when a maximum number of iterations is attained, or when the optimal fitness value converges.

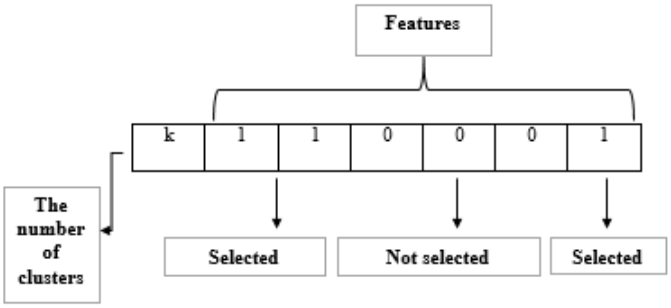

Figure 2. A representation of feature selection in the proposed algorithm

The answer is depicted in Figure 2, which also demonstrates how the feature selection and K value are encoded in binary string format. The K value is represented by the first portion of the string, while the feature selection is represented by the second part. The associated feature is selected when the number 1 is displayed and not selected when the number 0 is displayed.

Despite these good results in feature selection, BHHOA shares with other metaheuristic algorithms some of its limitations, namely stochastic variability and the risk of premature convergence in the search space. Its operation is also sensitive to control parameters, requiring fine-tuning. To address this, the experiments have been repeated, and adaptive or hybrid enhancements could be a direction for future work.

Regarding reproducibility, initialization was performed through the K-means algorithm, and the encoding of BHHOA agents was done as binary strings of feature subsets. Everything was treated continuously, and fitness was measured using the Calinski–Harabasz (CH) index. Each of the experiments was carried out 25 times using a static random seed.

5. Results and discussion

Five distinct classification datasets (*Absenteeism at work*, *Wholesale customers data*, *NPHA-doctor-visits*,*tripadvisor_review1*, and *Maternal Health Risk Data Set*) were subjected to K-means to validate the suggested approach. All of the datasets utilized were binary and were sourced from Blake and Merz (1998). A total of 20% of the data was allocated to the test group, while the remaining 80% was used for training.

Table 1. Description of the used datasets

Datasets	Cases	Dimensions
Data1 (Absenteeism at work)	740	20
Data2 (Wholesale customers data)	440	7
Data3 (NPHA-doctor-visits)	714	7
Data4 (tripadvisor_review1)	980	9
Data5 (Maternal Health Risk Data Set)	1014	6

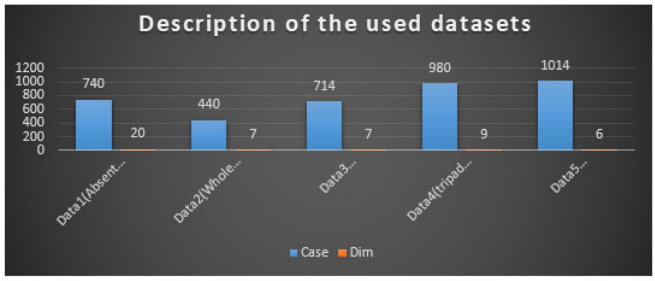

Figure 3. A description of the used datasets

After implementing the suggested algorithm (BHHO) to enhance the feature selection process in comparison to the average K-means, the findings indicated that the new algorithm excels in both performance and accuracy when identifying features that significantly influence modelling outcomes. Furthermore, a quantitative analysis comparing the proposed algorithm with traditional methods revealed a notable enhancement in performance metrics, including overall accuracy and a reduction in the number of selected features, all while maintaining the quality of the predictive models, as shown in Table 2.

Table 2. Description of the used BHHOA with implementation of K-means using the Calinski-Harabasz (CH) index

Datasets	GNDOA-FCM	FCM
Data1	0.4166	0.2689
Data2	0.8036	0.3283
Data3	0.3918	0.1308
Data4	0.5316	0.1857
Data5	0.6743	0.3752

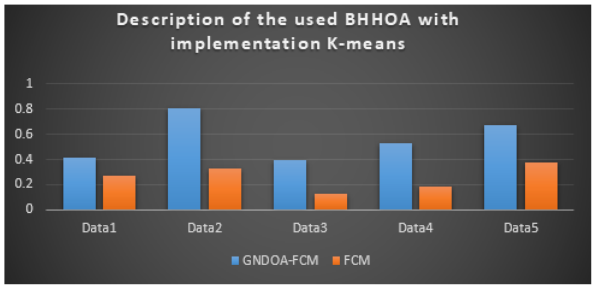

Figure 4. A representation of the used BHHOA with implementation of K-means

In Table 2, the proposed BHHOA with K-means implementation demonstrates superior efficiency across various datasets compared to traditional FCM, showcasing its robustness in handling diverse data distributions.

Table 3. Comparison of BHHOA - K-Means and K-Means algorithms in terms of feature selection

Datasets	BHHOA - K-Means	K-Means
Data1	11	20
Data2	1.6	7
Data3	3.6	14
Data4	2	9

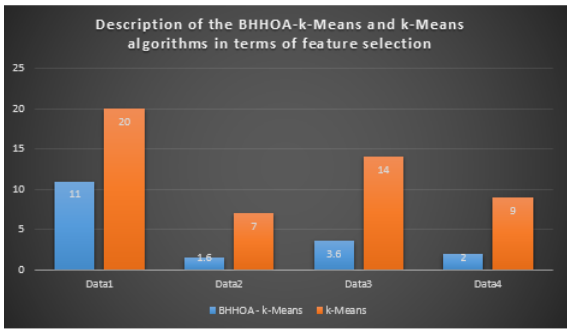

Figure 5. A representation compares the BHHOA-k-Means and k-Means algorithms in terms of feature selection

This advancement demonstrates that the proposed algorithm effectively balances model simplification with minimizing bias linked to unnecessary features. To validate the efficacy of the proposed algorithm, various evaluation metrics, such as classification accuracy and execution time, were employed. These metrics indicated that the algorithm decreases computational complexity and enhances efficiency relative to other approaches. These results support the notion that the proposed algorithm offers an innovative and effective solution to feature selection in practical applications.

6. Conclusions

The conducted research suggested a new approach that integrated the Binary Harris Hawks Optimization Algorithm (BHHOA) and K-means clustering to improve the classification of high-dimensional datasets. In this manner, BHHOA was utilized to achieve the most suitable dimensionality reduction and clustering results, especially after no justification was found for retaining unnecessary features within the datasets of interest.

Our approach proved to be more accurate and efficient than traditional methods across five datasets. The integration of BHHOA also demonstrated superior performance in feature selection, leading to more meaningful, non-overlapping cluster subdivisions. Thus, BHHOA showed its capability in handling complex, large-scale data.

This method provides a trade-off between computation time and clustering performance, offering a viable solution for clustering in high-dimensional spaces. Future research may involve applying the proposed BHHOA-based feature selection to real-world applications, such as medical image processing and remote sensing, to effectively address high-dimensional and noisy data, thereby enhancing clustering efficacy.

REFERENCES

1. B. Yang, et al., *Feature selection based on modified bat algorithm*, IEICE TRANSACTIONS on Information and Systems, vol. 100, no. 8, pp. 1860–1869, 2017.

2. M. Dong, Y. Wang, Y. Todo, and Y. Hua, "A novel feature selection strategy based on the Harris Hawks Optimization algorithm for the diagnosis of cervical cancer," *Electronics*, vol. 13, no. 13, p. 2554, 2024.
3. H. M. Alabool, D. Alarabiat, L. Abualigah, and A. A. Heidari, "Harris Hawks Optimization: a comprehensive review of recent variants and applications," *Neural Computing and Applications*, vol. 33, pp. 8939–8980, 2021.
4. A. Issa, Y. Ali, and T. Rashid, "An Efficient Hybrid Classification Approach for COVID-19 Based on Harris Hawks Optimization and Salp Swarm Optimization," *International Journal of Online and Biomedical Engineering (iJOE)*, vol. 18, no. 13, pp. 113–130, 2022.
5. A. A. Heidari, S. Mirjalili, H. Faris, I. Aljarah, M. Mafarja, and H. Chen, "Harris Hawks Optimization: Algorithm and Applications," *Future Generation Computer Systems*, vol. 97, pp. 849–872, 2019, doi:10.1016/j.future.2019.02.028.
6. B. Tripathy, et al., *Harris hawk optimization: a survey on variants and applications*, Computational Intelligence and Neuroscience, vol. 2022, no. 1, p. 2218594, 2022.
7. Y. Zhang, et al., *Boosted binary Harris hawks optimizer and feature selection*, Engineering with Computers, vol. 37, pp. 3741–3770, 2021.
8. K.P. Sinaga, and M.-S. Yang, *Unsupervised K-means clustering algorithm*, IEEE Access, vol. 8, pp. 80716–80727, 2020.
9. L. Li, J. Wang, and X. Li, *Efficiency analysis of machine learning intelligent investment based on K-means algorithm*, IEEE Access, vol. 8, pp. 147463–147470, 2020.
10. A. Bouguettaya, et al., *Efficient agglomerative hierarchical clustering*, Expert Systems with Applications, vol. 42, no. 5, pp. 2785–2797, 2015.
11. L. Abualigah, et al., *Hybrid Harris hawks optimization with differential evolution for data clustering*, in *Metaheuristics in machine learning: theory and applications*, Springer, pp. 267–299, 2021.
12. C. Qu, et al., *Harris Hawks optimization with information exchange*, Applied Mathematical Modelling, vol. 84, pp. 52–75, 2020.
13. A. S. Menesy, et al., *Developing and Applying Chaotic Harris Hawks Optimization Technique for Extracting Parameters of Several Proton Exchange Membrane Fuel Cell Stacks*, IEEE Access, vol. 8, pp. 1146–1159, 2020.
14. N.A. Golilarz, et al., *A New Automatic Method for Control Chart Patterns Recognition Based on ConvNet and Harris Hawks Meta Heuristic Optimization Algorithm*, IEEE Access, vol. 7, pp. 149398–149405, 2019.
15. O.M. Ismael, O.S. Qasim, and Z.Y. Algamal, *A new adaptive algorithm for v-support vector regression with feature selection using Harris hawks optimization algorithm*, in *Journal of Physics: Conference Series*, IOP Publishing, 2021.
16. Q. Fan, Z. Chen, and Z. Xia, *A novel quasi-reflected Harris hawks optimization algorithm for global optimization problems*, Soft Computing, 2020.
17. Jasim, M. D., Abduljabbar, A. S., and Dahash, N. M., "Finding a new parallel method of harmonic Runge-Kutta mean by using predictor corrector form," *Journal Name*, vol. X, no. Y, pp. ZZZ–ZZZ, Year.
18. O.S. Qasim, and Z.Y. Algamal, *A gray wolf algorithm for feature and parameter selection of support vector classification*, International Journal of Computing Science and Mathematics, vol. 13, no. 1, pp. 93–102, 2021.
19. S.G.M. Al-Kababchee, O.S. Qasim, and Z.Y. Algamal, *Improving penalized regression-based clustering model in big data*, in *Journal of Physics: Conference Series*, IOP Publishing, 2021.