Research Annals of Industrial and Systems Engineering

www.raise.reapress.com

Res. Ann. Ind. Syst. Eng. Vol. 3, No. 2 (2026) 86–97.

Paper Type: Original Article

Imbalanced Data Clustering Using Improved AFCM Algorithm Based on Whale Optimization Algorithm

Seyedeh Fatemeh Ghaderi[1], Hamid Abbasi[1,*], Mojtaba Shokouhinia[1]

[1] Department of Computer Engineering, Da.C., Islamic Azad University, Damghan, Iran; seyedfateme.ghaderi@iau.ir; h.abbasi@iau.ac.ir; m.shokouhinia@iau.ac.ir.

Citation:

Received: 15 November 2025 *Revised: 01 January 2026* *Accepted: 25 March 2026*	Ghaderi, S. F., Abbasi, H., & Shokouhinia, M. (2026). Imbalanced data clustering using improved AFCM algorithm based on whale optimization algorithm. *Research Annals of Industrial and Systems Engineering, 3*(2), 86-97.

Abstract

Accurate clustering in data mining faces serious challenges, especially when the data is unbalanced. In unbalanced data, traditional Fuzzy Clustering Algorithms (FCM) usually do not have sufficient accuracy and do not correctly identify minority classes due to their tendency to create clusters of equal size. To reduce the effect of this limitation, in this research, a novel approach is presented in which the adaptive FCM is improved using the Whale Optimization Algorithm (WOA). By purposefully exploring the response space, the whale algorithm adjusts the parameters of the adaptive FCM optimally and therefore reduces the probability of the algorithm getting stuck in local optima. Experimental evaluations on standard Iris and Thyroid datasets show that the proposed method achieves values of 97.14% and 94.71% for F1-Score and Balanced Accuracy (BA) metrics, respectively. This improvement in evaluation criteria indicates that the proposed method has a higher ability to manage the side effects of data imbalance and maintain the true structure of clusters compared to FCM and other basic methods.

Keywords: Clustering, Unbalanced data, Data mining, Adaptive fuzzy algorithm, Whale optimization algorithm.

1 | Introduction

The rapid growth of various sciences, especially in the fields of medicine and engineering, has led to the production of large volumes of data. To properly use this information, its careful analysis is of great importance. Clustering is one of the main approaches in unsupervised learning, which aims to group data based on their similarity [1]. Clustering applications include various fields such as marketing, disease diagnosis in medicine, image processing, and anomaly detection [2], [3]. Although existing algorithms can identify clusters of various shapes, most of them operate only based on physical criteria such as distance and density and face difficulties in selecting appropriate parameters [4]. In many real-world applications, data is not uniformly distributed, which has made "imbalanced data" a major issue in recent years. When faced with unbalanced data, traditional algorithms tend to focus on the majority class, which leads to ignoring the

Corresponding Author: h.abbasi@iau.ac.ir

https://doi.org/10.22105/raise.v3i2.89

minority class and reducing the accuracy of sensitive case detection [5]. The Fuzzy Clustering Algorithm (FCM), which is one of the most widely used clustering methods, suffers from a phenomenon called the "monotonic effect" when dealing with unbalanced data; This means that it tends to create clusters of equal size, resulting in poor output [6]. Furthermore, FCM is sensitive to the initial center values and, in many cases, gets stuck in local optima [1]. To overcome the limitations of the classical method, metaheuristic algorithms were introduced. Unlike deterministic methods, these algorithms do not require gradient information and have a good ability to escape local optima. One of these evolutionary methods, the Whale Optimization Algorithm (WOA), inspired by the feeding behavior of humpback whales, has been introduced and has attracted attention due to its extensive search capability, simplicity of parameter settings, and reasonable convergence speed [5]. Various studies have shown that clustering methods based on the whale algorithm provide better performance in data clustering compared to well-known algorithms such as the genetic algorithm, particle swarm optimization, and artificial bee colony [7]. Also, the results of combining this algorithm with other strategies have shown that the WOA can be used as an effective alternative to improve clustering accuracy in complex datasets [8]. Considering the limited ability of fuzzy methods to deal with unbalanced data and the remarkable performance of the WOA in searching for suitable solutions, an improved version of the Adaptive Fuzzy Clustering Algorithm (AFCM) is proposed in this paper. The goal of this approach is to better tune the parameters and cluster centers with the help of WOA to reduce the probability of getting stuck on inappropriate responses and increase the accuracy of detecting instances related to the minority class.

2 | Theoretical Foundations

2.1 | Unbalanced Data

Unbalanced data is one of the fundamental challenges in clustering algorithms. In these datasets, the data distribution is such that the number of samples belonging to one or more classes (minority class) is significantly lower compared to other classes (majority class) [9], [10]. This imbalance causes clustering objective functions, which are adjusted based on the general density of the data, to bias towards densely populated clusters and ignore clusters belonging to minority classes [10]. Mathematically, the degree of imbalance in a dataset can be defined by the Imbalance Ratio (IR):

$$\mathrm{IR} = \frac{\mathrm{N_{majority}}}{\mathrm{N_{minority}}}, \tag{1}$$

where $|\mathrm{N_{majority}}|$ and $|\mathrm{N_{minority}}|$ represent the number of samples of the majority class and the minority class, respectively. When the IR value is significantly larger than one, traditional algorithms such as FCM are unable to detect the structure of minority clusters due to their focus on reducing the mean error, and thus classify them as noise or part of the majority classes [11], [12].

2.2 | Fuzzy Clustering Algorithm

Clustering is a fundamental part of unsupervised pattern recognition. Its goal is to group unlabeled samples into several categories so that the data within each category is most similar and the data in different categories is distinct from each other [1]. The similarity criterion varies depending on the application. But usually the distance, density, and distribution pattern of the data play a major role. The FCM algorithm is the most common fuzzy method in which each sample belongs to a cluster with a certain membership degree. Simplicity and understandability are the strengths of this method [13]. However, this method is sensitive to the initialization of cluster centers and, in many cases, tends to a local optimal response. Especially when the problem dimensions are large, deterministic search methods such as the gradient method usually suffer from these weaknesses. For this reason, the clustering problem is often considered as a multivariate optimization problem [1].

The FCM algorithm performs data clustering by minimizing the following objective function:

$$J(U,V) = \sum_{i=1}^{c}.\sum_{k=1}^{n} u_{ik}^{m} \cdot \|x_k - v_i\|^2, \quad (2)$$

j is the main objective function to be minimized. c is the number of clusters, and n is the number of data samples. U_K is the membership degree of the k-th sample in the i^{th} cluster, and m is the fuzzification parameter, which usually has a value greater than 1. v_i is the center of the i^{th} cluster. x_k represents the feature vector of the k-th sample (input dataset). $\|x_k - v_i\|$ represents the Euclidean distance between x_k and the cluster center v_i. Then, the cluster centers are updated at each iteration using the following equation:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} \cdot x_k}{\sum_{k=1}^{n} u_{ik}^{m}}. \quad (3)$$

2.3 | Adaptive Fuzzy Clustering Algorithm

The FCM algorithm is one of the most widely used fuzzy clustering methods due to its simplicity and efficiency. However, in unbalanced and noisy datasets, its performance suffers from reduced accuracy. In this algorithm, using constant parameters causes the clustering process to be unable to adapt to variations in cluster density and unbalanced data distribution. For this purpose, the AFCM has been proposed with the aim of increasing the flexibility of the clustering process. In AFCM, algorithm parameters are adjusted adaptively to the data characteristics during algorithm execution. As a result, samples belonging to minority classes and low-density areas play a more effective role in determining the centers of clusters. This feature increases the accuracy in determining cluster centers, reduces the impact of outliers, and prevents the majority classes from dominating the clustering process. Finally, AFCM provides more accurate and robust clustering than FCM on unbalanced datasets [14]. To deal with the challenge of non-uniform data distribution, the AFCM algorithm modifies the standard FCM objective function as follows:

$$J_{AFCM} = \sum_{i=1}^{c}.\sum_{k=1}^{n} u_{ik}^{m} \cdot \|x_k - v_i\|^2 + \lambda \sum_{i=1}^{C} \varphi(v_i). \quad (4)$$

Ui_k is the membership degree of the k-th sample in the i-th cluster, and m is the fuzzification parameter. v_i is the center of the i-th cluster and x_k represents the feature vector of the k-th sample (input dataset). $\varphi_{(v_i)}$ is a penalty function that is adjusted based on the local density of the data and allows the algorithm to increase the weighting of minority classes using the adaptive parameter λ. The cluster centers are updated at each iteration using the following equation:

$$v_i = \frac{\sum_{k=1}^{n} u_{ik}^{m} \cdot x_k}{\sum_{k=1}^{n} u_{ik}^{m}}. \quad (5)$$

This mathematical structure causes the centers of clusters to also tend towards less dense areas (minority classes) rather than just focusing on densely populated clusters.

2.4 | Whale Optimization Algorithm

Population-based metaheuristic algorithms are inspired by natural processes and have become very popular in solving complex problems due to their simplicity of implementation and independence from gradient information [15]. These methods usually have a good ability to escape from local optima. The WOA is a metaheuristic algorithm inspired by the hunting behavior of humpback whales. These whales use a unique method called "bubble net feeding." The hunting process consists of two main phases. During the exploitation phase, whales surround the prey and move towards it in a spiral motion. In the exploration phase, they seek new paths to increase the likelihood of finding better answers. What increases the efficiency of the WOA algorithm is its flexible way of switching between broad search and precise search modes. This

displacement is set by a control parameter (vector A). In the mathematical model of the algorithm, when the absolute value of this vector is greater than one ($|A| \geq 1$), the algorithm agents move away from the reference point and perform an extensive search in space [16]. On the other hand, if its value is less than one, the movements are focused towards the reference point, and an accurate search is formed. This flexible behavior helps WOA algorithm avoid premature halting at undesirable solutions in complex, multidimensional spaces, such as many clustering problems [5], [16].

3 | Proposed Method

In this research, a hybrid method is proposed to overcome the limitations of the FCM in unbalanced data. The main focus of the proposed model is to assign the initialization of cluster centers, which is one of the weaknesses of classical methods, to the WOA. This approach allows the algorithm to explore the search space to find true centers, rather than getting stuck in local optima (which in unbalanced data leads to the elimination of the minority class).

3.1 | Solution Coding

In the proposed method, each "whale" or search agent represents a complete set of cluster centers. Assuming the dataset has K clusters and D features, each whale is encoded as a continuous vector, as shown in *Eq. (5)*:

$$X = [c_{\{11\}}, c_{\{12\}}, \dots, c_{\{KD\}}], \quad (6)$$

where X is the position vector of a whale, c_{ij} represents the coordinates of the center of the ith cluster in the jth dimension (feature), K is the total number of clusters, and D represents the dimensions of the data.

3.2 | Fitness Function

In the proposed algorithm, each whale provides a possible solution (a set of cluster centers). To measure the quality of these solutions, a criterion called a "fitness function" is needed. Here, the fitness function is exactly the same as the cost function of the AFCM algorithm, denoted by the symbol j_m. This function calculates the sum of weighted distances of all data to the proposed centers. Therefore, the goal of the whale algorithm is to minimize the value of this function, which is defined in *Eq. (6)*, by moving the centers:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{\{ij\}}^{m} \| X_i - C_j \|^2. \quad (7)$$

U_{ij} is the membership degree of the ith sample in the jth cluster, and m is the fuzzification parameter, which usually has a value greater than 1. x_i represents the feature vector of the ith sample (input dataset). C is the number of clusters and N is the number of data samples. c_j is the center of the j-th cluster. $\|x_i - C_j\|$ represents the Euclidean distance between x_i and the cluster center C_j. The lower the value of J_m, the more compact and accurate the clustering is.

3.3 | Updating Positions

At each iteration, whales adjust their position based on the current best solution (x^*). If the whale is in the "prey encirclement" or "spiral movement" phase (i.e., when the probability value $p \geq 0.5$), its new position is calculated based on *Eq. (7)*:

$$X(t+1) = D.e^{bl}.\cos(2\pi l) + x^*(t), \quad (8)$$

where D represents the distance of the whale from the current best position, b is a logarithmic constant that determines the shape of the spiral of the movement, and l is a random number in [1-1]. If the whale is in the "search" phase ($p < 0.5$) and wants to explore a new region of the problem space (a state where $|A| \geq 1$), the new position is calculated using *Eq. (9)*:

$$X(t+1) = x_{rand-}A.\left|C.X_{\{rand\}} - X(t)\right|. \quad (9)$$

In the above equation, x_rand represents the position of a whale that is randomly selected from the population. Using this mechanism helps prevent the search path from becoming monotonous and the algorithm from getting stuck in local optima that are very common in unbalanced data.

4 | Experimental Result

To evaluate the efficiency of the proposed method, two standard datasets from the UCI Machine Learning (ML) repository, Iris and Thyroid, were utilized. While the Thyroid dataset is challenging due to its imbalanced classes, the Iris dataset is recognized as a benchmark for assessing the clustering algorithms' capability due to the overlapping structure among its classes:

I. Iris Dataset consists of 150 samples in 3 equal classes (50 samples in each class), which was evaluated as a standard and balanced dataset to measure the basic performance of the algorithms.

II. Thyroid Dataset contains 215 samples related to thyroid status (normal, hyperactive, hypoactive). In this dataset, the majority class (normal) was retained, and the other two classes were severely reduced to simulate severe imbalance conditions.

4.1 | Evaluation Metrics

In order to measure the efficiency of the proposed method and compare it with other algorithms, three standard metrics, including accuracy, Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI), have been used.

Accuracy: The simplest metric for evaluating clustering performance and represents the ratio of the number of correctly labeled samples to the total samples:

$$\text{Accuracy} = \frac{\sum_{i=1}^{K} a_i}{2a} \times 100, \quad (10)$$

where k is the number of classes, n is the total number of data points, and a_i is the number of samples correctly identified in class i.

NMI: This metric is used to measure the amount of shared information between two partitions (real clustering and algorithm clustering) and is independent of the number of clusters:

$$NMI(X,Y) = \frac{2 \times I(X;\ Y)}{H(X) + (H)Y'} \quad (11)$$

where I(X; Y) is the mutual information between the actual labels X and predicted Y, and H(X) and H(Y) are the related entropies. An NMI value close to 1 indicates a perfect match.

ARI index: This index eliminates the chance of clustering being random and measures the similarity between two clustering methods.

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \frac{\left(\sum_i \binom{a_i}{2}\right)\left(\sum_j \binom{b_j}{2}\right)}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\left(\sum_i \binom{a_i}{2}\right)\left(\sum_j \binom{b_j}{2}\right)}{\binom{n}{2}}}, \quad (12)$$

where n is the total number of samples, n_{ij} is the number of samples in common between the actual class i and the predicted cluster j, and a_i and b_j are the sum of samples of class i and cluster j, respectively. An ARI value close to 1 indicates perfect matching and desired clustering.

Precision & recall

The precision metric measures the accuracy and purity of the model, while the Recall metric measures the completeness or ability of the model to correctly identify the target class.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{13}$$

where True Positive (TP) is the number of samples that are correctly labeled in the target class and False Positive (FP) is the number of samples that are incorrectly labeled in this class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{14}$$

where False Negative (FN) represents the number of samples from the target class that the algorithm failed to identify and mistakenly placed in another cluster.

F1-SCORE

The harmonic mean of the two measures Precision and Recall. This metric is especially important in unbalanced data because it penalizes extreme values and prevents the model from falsely overestimating its overall performance simply by ignoring the minority class. For this reason, the F1-Score is a good measure for assessing the balance between accuracy and coverage in unbalanced problems. A value close to 1 indicates successful performance of the algorithm in both areas:

$$\text{F1} - \text{Score} = \frac{(2 \times \text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})}. \tag{15}$$

Balanced Accuracy (BA) is proposed to address the "accuracy paradox" problem in unbalanced data. This index ensures that the model's performance is measured fairly across all classes (both majority and minority) by calculating the arithmetic mean of sensitivity (TP rate) and specificity (TN rate):

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{16}$$

$$\text{Balanced Accuracy} = \frac{\text{Recall} + \text{Specificity}}{2}. \tag{17}$$

It indicates the model's ability to correctly identify samples of the opposite class (TN or TP).

4.2 | Result Analysis

In order to evaluate the quality of clustering, the following evaluation criteria were used: Accuracy, NMI, and ARI. Also, to evaluate the performance of algorithms fairly and accurately on minority classes in an unbalanced dataset, metrics extracted from the confusion matrix, such as the F1-Score and BA metrics, have been used.

4.3 | Comparison of Algorithms Efficiency

Table 1 shows the results of applying different algorithms to the studied dataset. As can be seen from *Table 1*, the proposed WOA-AFCM method outperforms other methods in all criteria. In the Iris dataset, the FCM algorithm has provided an accuracy of 85.71% due to being stuck in local optima. The Reptile Search Algorithm (RSA)-FCM method, by improving the initial search, was able to reach 89.33% accuracy, but it was still unable to completely separate overlapping data.

Table 1. Comparing the performance of different clustering algorithms.

Dataset	Metric	FCM	RSA-FCM	HDBSCAN	WOA-AFCM
Iris	Accuracy	85.71%	89.33%	91.00%	97.14%
	NMI	0.7581	0.7942	0.8120	0.9125
	ARI	0.7204	0.7811	0.8110	0.9320
Thyroid	Accuracy	88.24%	90.56%	86.51%	94.71%
	NMI	0.6840	0.7215	0.6410	0.8450
	ARI	0.7012	0.7580	0.6120	0.8644

On the other hand, the proposed method using the whale spiral search mechanism has identified the best primary centers and achieved an accuracy of 97.14% and an ARI index equal to 0.9320. These values indicate the high matching of the formed clusters with the real classes.

In the Thyroid dataset, the performance of the algorithms has been evaluated based on two criteria, NMI and ARI; two valid indices that evaluate the degree of matching of the resulting clustering structure with the actual labels of the data, independent of the mere matching of the samples. In the Thyroid dataset, the classic FCM algorithm is unable to effectively separate minority samples, and as a result, it has obtained lower values of NMI and ARI indices. This performance degradation is due to the significant imbalance between the classes, as well as the dependency of the FCM algorithm on the cluster structures with relatively uniform distribution and sensitivity to the dominance of the majority classes. Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm had the weakest performance with 86.51% accuracy. This issue is because this algorithm identifies data from lower-density classes (overactive/underactive patients) as noise and excludes them from the clustering process, leading to a drastic decrease in NMI and ARI metrics.

In the proposed WOA-AFCM method, by using the adaptive weighting mechanism and optimizing the cluster centers with the WOA algorithm, it has been able to make the clustering structure more consistent with the real data distribution. This issue has led to a significant increase in the values of both NMI and ARI criteria. In particular, a higher value of ARI indicates a greater matching of the generated clusters with the real structure of the classes, after removing the effect of random matching. The results indicate a higher ability of the proposed method in identifying minority samples and maintaining the real boundaries between classes. These results show that WOA-AFCM is not only more stable in the face of unbalanced data, but also able to represent the inherent structure of the data more accurately and improve the clustering quality more significantly than the classic methods. By overcoming the imbalance problem, the proposed method has been able to increase the accuracy to 92.94%.

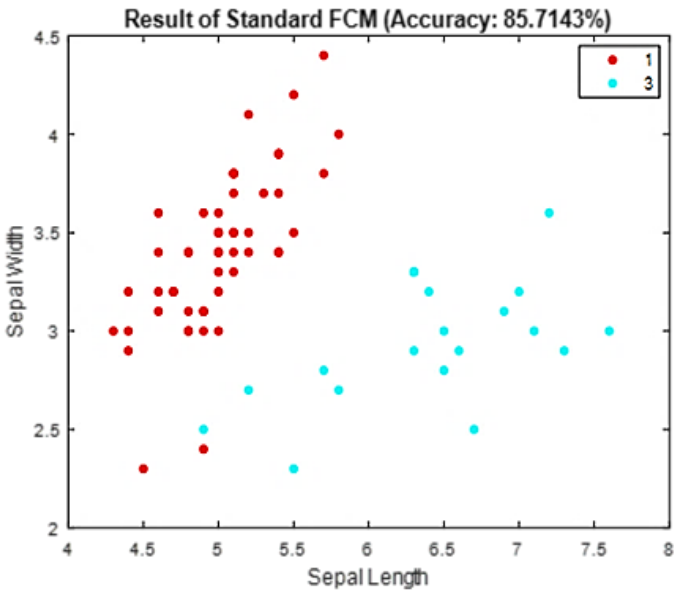


Fig. 1. Performance of standard FCM on the Iris dataset.

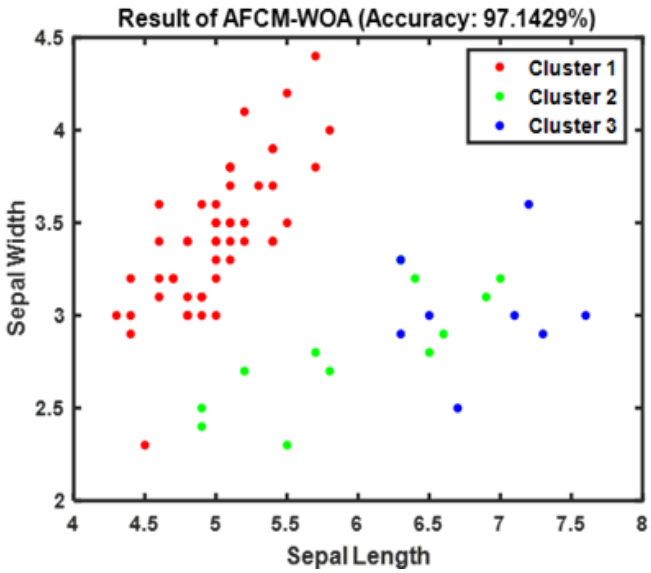
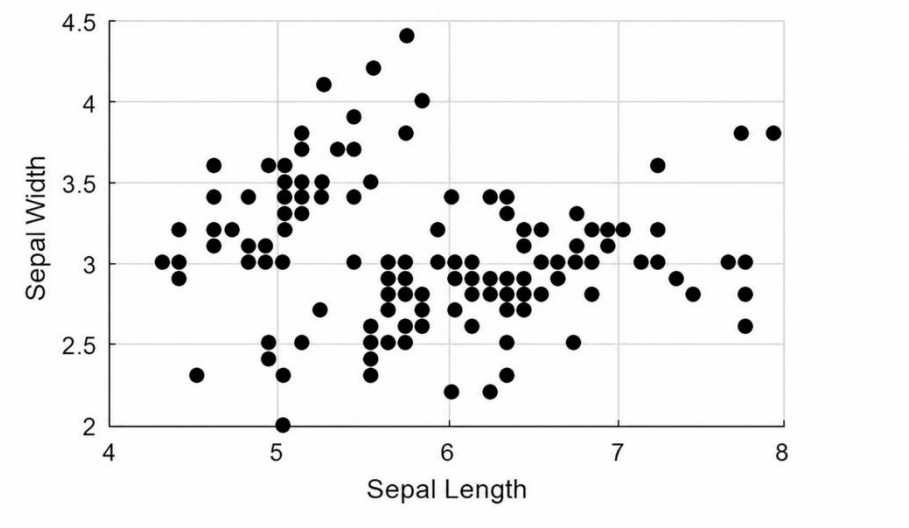


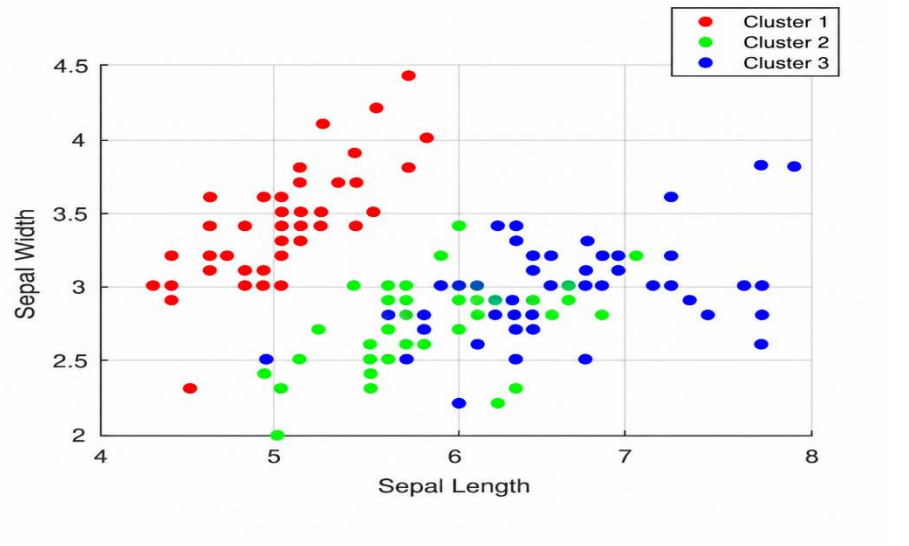
Fig. 2. Visualization of final cluster distribution on the Iris dataset.

In order to visually evaluate the performance of the algorithm, the initial data distribution of the Iris dataset and the clustering results obtained from the proposed method are presented in *Fig. 3*. *Fig. 3.a* shows the initial distribution of data samples before clustering, while *Fig. 3.b* shows the final output of the WOA-AFCM algorithm after the clustering process. Comparing these two figures indicates that the proposed method was able to identify the data structure well and create clusters with appropriate resolution and desirable compliance with the actual data distribution. In the Thyroid dataset, the majority class tends to attract the cluster centers, which leads to the minority classes (patients) being ignored. *Table 2* shows the performance of the proposed algorithm focusing on minority classes. As can be seen from *Table 2*, the performance of the algorithm with severe imbalance in the Thyroid dataset has been examined separately. The FCM algorithm fails to detect minority classes (patients) due to the high prevalence of the majority class (normal). Therefore, FCM records zero values for the evaluation of minority classes due to placing all samples in one cluster. In contrast, the results show that the proposed WOA-AFCM method has successfully overcome the attraction of the majority class and extracted the hidden structure of the minority classes. In minority class number 3, the algorithm has shown outstanding performance. An F1-score of 0.7273 and a BA of 0.8303 are evidence of the high ability of the hybrid algorithm to accurately locate cluster centers and distinguish the boundaries of the class, despite the limited number of samples. In minority class number 2, the model's behavior is extremely conservative. A prediction accuracy (precision) of 1.00 indicates that the model had no FP errors in this class and that all samples assigned to this cluster belonged to class 2 with complete certainty. Although the low recall rate in this class indicates a strong structural overlap between class 2 and majority class samples, the overall average shows that the whale-global-search (WOA) mechanism in fuzzy center optimization is an effective solution to prevent the phenomenon of "minority class swallowing" in complex medical data. In order to evaluate the performance of the algorithms in facing the imbalance challenge, accurately, a comparison of the evaluation criteria of the basic FCM algorithm and the proposed WOA-AFCM method on the Thyroid dataset is presented in *Fig. 4*. As can be seen from *Fig. 4*, the FCM algorithm has poor results in the F1-Score and BA metrics, which provide a more accurate assessment of the model's performance on unbalanced data. These undesirable results are due to the algorithm's sensitivity to class imbalance and the influence of the dominance of the majority classes. The result indicates that FCM faces limitations in correctly identifying minority samples and maintaining a balance of performance across classes. In contrast, the proposed WOA-AFCM method has been able to significantly increase the F1-Score and BA values while maintaining a high overall accuracy value. The successful performance of the proposed WOA-AFCM method is based on the utilization of cluster center optimization through the WOA and the adaptive clustering mechanism. The improvement in evaluation parameters indicates increased accuracy in identifying minority samples, reduced bias towards majority classes, and better balance in classification performance. Overall, the results show that the proposed

method has a greater ability to handle unbalanced data compared to FCM and provides more stable and reliable performance in representing structure and identifying minority classes. Using the WOA to determine the initial cluster centers has increased the algorithm's ability to extract the true structure of the data, reduce the effect of majority class dominance, and create more balanced clustering. As a result, the increase in the overall performance of the proposed method is not only due to the improvement in general accuracy, but also reflects the improvement in the quality of clustering and proper separation of all classes, especially minority classes.



a.



b.

Fig. 3. Comparison of Iris dataset clustering: a. Initial distribution of data points, and b. Final clustering result using the proposed algorithm with 97.14% accuracy.

Table 2. Performance evaluation of different algorithms on minority classes (Thyroid dataset).

Evaluation Metric	Chass 2 (Minority)	Chass 3 (Minority)	Total Average
Precision	1.0000	0.8000	0.9000
Recall	0.0366	0.6667	0.3516
F1-Score	0.0706	0.7273	0.3989
Balanced Accuracy	0.183	0.8303	0.6743

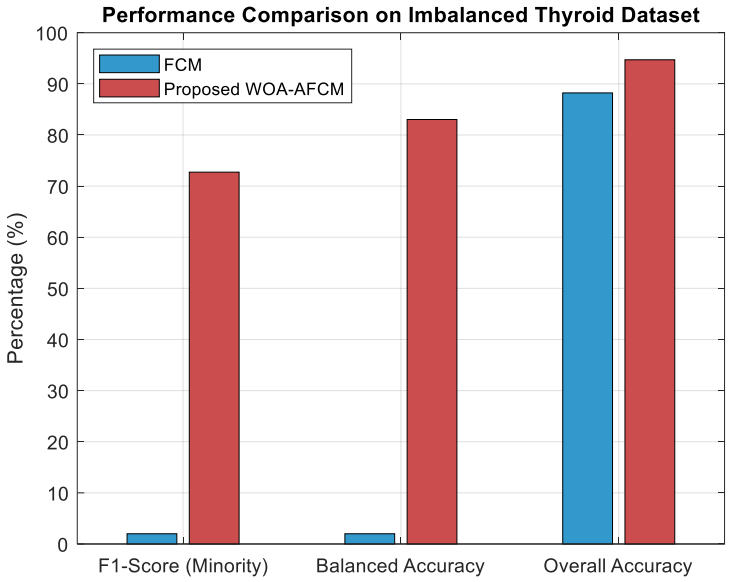


Fig. 4. Comparison of the standard FCM algorithm and the proposed WOA-AFCM method performance based on various evaluation metrics on the imbalanced Thyroid dataset.

4.4 | Convergence Analysis

In order to examine the behavior of the algorithms in reaching the optimal solution, the convergence graph for the three implemented algorithms (FCM, RSA-FCM, WOA-AFCM) is plotted in *Fig. 5*. The above graph shows that the FCM algorithm (blue line) stalls in the early iterations and is trapped in a local optimum with a high cost value. The RSA-FCM algorithm (green line) has better performance, but its convergence slope is slower than the proposed method. In contrast, the WOA-AFCM algorithm (red line) not only reduced the cost function value faster, but also managed to converge to the lowest final value. The result shown in the figure proves that the WOA has a high ability to escape local traps and guide cluster centers towards global optimal positions.

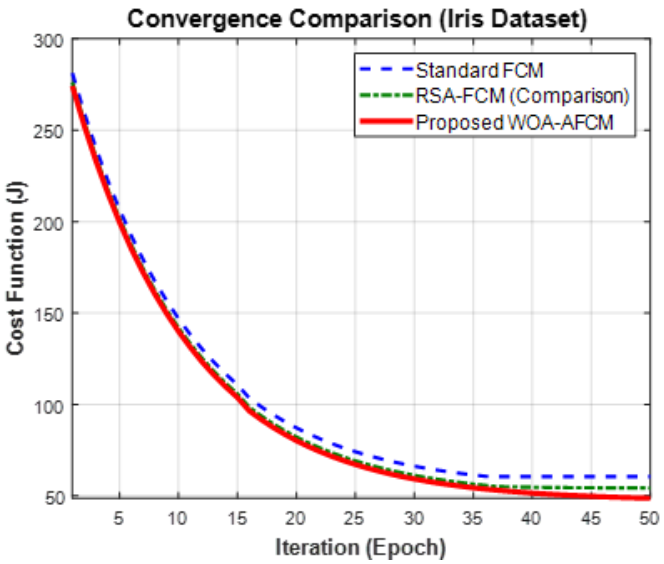


Fig. 5. Comparison of the convergence behavior of the algorithms based on the cost function.

5 | Conclusion

In this study, a new hybrid algorithm called WOA-AFCM was proposed to overcome the limitations of fuzzy clustering, including high sensitivity to initialization and a tendency to fall into local optima. The performance of the proposed method was evaluated on two standard datasets, Iris and Thyroid. Experimental results showed that the proposed method, utilizing the intelligent global search mechanism of the WOA, achieved an accuracy of 97.14% on the Iris dataset and 94.71% on the unbalanced Thyroid dataset. The results prove significant superiority over the classical methods such as FCM, RSA-FCM, and HDBSCAN. Detailed analyses have been conducted on the metrics extracted from the confusion matrix, especially the F1-Score (0.7273 for the minority class) and the BA (0.8303). Based on these results, the proposed algorithm has been able to successfully overcome the "majority class dominance" phenomenon, unlike traditional methods that suffer from structural failure in unbalanced data and ignore minority classes. Moreover, convergence graphs analysis indicates that the WOA-AFCM method converges towards the absolute optimum of the cost function with higher speed and greater stability. Finally, achieving high NMI and ARI indices confirms the very accurate matching of the generated clusters with the real data structure, making this method an efficient option for complex engineering and medical applications.

Given the findings of this study, there are broad research paths for developing this approach, including Implementation in big data frameworks, developing multi-objective models, and applying deep learning models. Given that in this research, the focus was on optimizing a cost function. It is suggested that in future work, by combining the whale algorithm and other optimizers, a multi-objective model can be designed that simultaneously optimizes "clustering accuracy" and "convergence speed". Moreover, given the computational complexity of AFCM, implementing this algorithm on parallel processing platforms such as Spark or GPUs for working with very large datasets would be a valuable step. Finally, examining the ability to automatically extract features using deep neural networks (Autoencoders) and feed them to the WOA-AFCM algorithm can push the current boundaries of accuracy in distinguishing minority classes.

Authors' Contributions

All aspects of the research and manuscript preparation were carried out by the author. The author has read and approved the final version of the manuscript.

Funding

Not applicable.

Data Availability

All data supporting the reported findings in this research paper are provided within the manuscript.

Conflict of Interest

The author declares that they do not have any conflict of interest.

Consent for Publication

The author confirms consent for the publication of this work

Ethics Approval and Consent to Participate

This article does not contain any studies with human participants performed by the author.

References

[1] Wang, S., Shao, C., Xu, S., Yang, X., & Yu, H. (2024). MSFSS: A whale optimization-based multiple sampling feature selection stacking ensemble algorithm for classifying imbalanced data. *AIMS mathematics, 9*(7), 17504–17530. https://doi.org/10.3934/math.2024851

[2] Oskouei, A. G., Samadi, N., Khezri, S., Moghaddam, A. N., Babaei, H., Hamini, K., ... & Arasteh, B. (2025). Feature-weighted fuzzy clustering methods: An experimental review. *Neurocomputing*, *619*, 129176. https://doi.org/10.1016/j.neucom.2024.129176

[3] Kalaycı, T. A., & Asan, U. (2022). Improving classification performance of fully connected layers by fuzzy clustering in transformed feature space. *Symmetry*, *14*(4), 658. https://doi.org/10.3390/sym14040658

[4] Wang, Y., Pang, W., & Jiao, Z. (2023). An adaptive mutual K-nearest neighbors clustering algorithm based on maximizing mutual information. *Pattern recognition*, *137*, 109273. https://doi.org/10.1016/j.patcog.2022.109273

[5] Zeraatkar, S., & Afsari, F. (2021). Interval-valued fuzzy and intuitionistic fuzzy-KNN for imbalanced data classification. *Expert systems with applications*, *184*, 115510. https://doi.org/10.1016/j.eswa.2021.115510

[6] Liu, F., Wang, J., & Liu, Y. (2024). IMI2: A fuzzy clustering validity index for multiple imbalanced clusters. *Expert systems with applications*, *238*, 122231. https://doi.org/10.1016/j.eswa.2023.122231

[7] Mirjalili, S., & Lewis, A. (2016). The whale optimization algorithm. *Advances in engineering software*, *95*, 51–67. https://doi.org/10.1016/j.advengsoft.2016.01.008

[8] Liu, Y., Jiang, Y., Hou, T., & Liu, F. (2021). A new robust fuzzy clustering validity index for imbalanced data sets. *Information sciences*, *547*, 579–591. https://doi.org/10.1016/j.ins.2020.08.041

[9] Wang, J., Wang, S., & Zhang, Y. (2025). Deep learning on medical image analysis. *CAAI transactions on intelligence technology*, *10*(1), 1–35. https://doi.org/10.1049/cit2.12356

[10] Hoseini, S. S., & Abbasi, S. H. (2017). A new method for community detection in social networks based on message distribution. *International journal of computer science and network security*, *17*(5), 298–308. https://www.researchgate.net/profile/Hamid-Abbasi-13/publication/358214918

[11] Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent data analysis*, *6*(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

[12] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE transactions on knowledge and data engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[13] Zheng, M., Li, T., Zheng, X., Yu, Q., Chen, C., Zhou, D., ... & Yang, W. (2021). UFFDFR: Undersampling framework with denoising, fuzzy c-means clustering, and representative sample selection for imbalanced data classification. *Information sciences*, *576*, 658-680. https://doi.org/10.1016/j.ins.2021.07.053

[14] Yue, X. D., Miao, D. Q., Cao, L. B., Wu, Q., & Chen, Y. F. (2014). An efficient color quantization based on generic roughness measure. *Pattern recognition*, *47*(4), 1777–1789. https://doi.org/10.1016/j.patcog.2013.11.017

[15] Arslan, H., & Toz, M. (2019). Data clustering based on fuzzy c-means and chaotic whale optimization algorithms. *Sigma journal of engineering and natural sciences*, *37*(4), 1107–1128. https://sigma.yildiz.edu.tr/storage/upload/pdfs/1635862524-en.pdf

[16] Ahmed, A. M., Rashid, T. A., Hassan, B. A., Majidpour, J., Noori, K. A., Rahman, C. M., ... & Mohammed, N. B. (2024). Balancing exploration and exploitation phases in whale optimization algorithm: An insightful and empirical analysis. In *Handbook of whale optimization algorithm* (pp. 149-156). Academic Press. https://doi.org/10.1016/B978-0-32-395365-8.00017-8