




# Journal of Soft Computing and Artificial Intelligence

Journal homepage: <https://dergipark.org.tr/en/pub/jscai>

International  
Open Access 

Volume 05  
Issue 02

December, 2024

## Research Article

### Artificial Algae Algorithm for Clustering of Benchmark Datasets

Sahar Rashedi<sup>1</sup> , Muhammed Eshaq Rashedi<sup>2</sup> , Murat Karakoyun<sup>3</sup> 

<sup>1,3</sup> Department of Computer Engineering, Necmettin Erbakan University, Konya, Turkey

<sup>2</sup> Department of Mathematics, Faculty of Science and Letter, Kafkas University, Kars, Turkey

#### ARTICLE INFO

##### Article history:

Received **October 2, 2024**

Revised **October 21, 2024**

Accepted **October 31, 2024**

##### Keywords:

Artificial Algae Algorithm  
Clustering  
Medical Datasets  
Optimization

#### ABSTRACT

The best solution found for a problem under specific circumstances is called optimization. Algorithms for optimization can make the best use of the information at their disposal. Numerous optimization algorithms have been created thus far by researchers, and most of these algorithms are based on the characteristics of naturally occurring biological organisms. Optimization algorithms have proven to be highly effective in numerous fields, including finance, engineering, and medical. Apart from these applications, they have also been employed in data mining techniques including clustering and classification. In many different domains, the clustering method is widely applied. Finding the optimum cluster centers is the most crucial step in the clustering process. In this study, the Artificial Algae Algorithm (AAA) is used to perform the clustering procedure by using 12 datasets that were taken from the UCI Machine Learning Repository. For every dataset, the squared distance values between the cluster centers and the data were computed in order to assess the effectiveness of AAA. The study evaluated AAA's performance against that of the ALO, DEA, MFO, PSO, TSA, and WOA algorithms. According to the experimental results, AAA took the first place by obtaining the best average values (0.72 for f1-score, 0.75 for sensitivity and 0.88 for specificity) in all three metrics, clearly demonstrating its success in the clustering problem.

## 1. Introduction

One use of data mining is the clustering issue, which involves organizing the items in a dataset based on similarities (or differences). Even in cases when the group to which the data belong is unknown, clustering algorithms aid in breaking the data down into subsets based on shared characteristics [1], [2]. Clustering is mostly used to group things based on shared characteristics. When grouping data into clusters, the goal is to make sure that members of the same group are comparable to one another, while members of different groups are placed in different groups. The goal is for individuals within a cluster to be close to one another, whereas individuals outside

of a cluster are farther apart [3]. Clustering is performed to the population when the grouping of the variables in the dataset is unclear. This population's  $n$  data samples are examined across  $p$  variables. People that share similar traits are grouped together and divided into clusters throughout the clustering process. The process of clustering makes it possible to aggregate observational data with little loss [4]. When performing cluster analysis, the first step is to select a similarity or distance criterion. Then, it must be decided which clustering technique will be used. The type of method for clustering which is utilized for the chosen approach is chosen in the following phase, and the number of clusters to be employed in interpreting the cluster results is decided in the last

\* Corresponding author

e-mail: [mkarakoyun@erbakan.edu.tr](mailto:mkarakoyun@erbakan.edu.tr)

DOI: 10.55195/jscai.1560068

step [5]. The Euclidean distance was used in this work to determine the separation between each data instance and the cluster center to which it belonged [6]. Numerous methods exist in the literature to address the clustering problem, as the text mentions. That means there are two main types of clustering algorithms: hierarchical and non-hierarchical. The number of clusters does not have to be predetermined when using hierarchical clustering techniques. After the clustering procedure is finished, the number of clusters is established using these methods. On the other hand, the number of clusters needed to complete the clustering process is necessary for non-hierarchical clustering algorithms. In conclusion, the division of  $n$  data instances into  $k$  clusters is the aim of non-hierarchical clustering. According to Boushaki, S. I., Kamel, N., and Bendjeghaba, the time complexity of hierarchical clustering methods is  $n^2$ , whereas non-hierarchical clustering methods have  $n$  complexity [7].

Optimization is the process of producing the most efficient result with limited resources under expected conditions. Due to the complexity of the problems, it is difficult to estimate all possible combinations for this generated result. Mathematical model approaches simplify the problems to solve these problems. They also make assumptions about the results to reduce the search space instead of all possible combinations. There can be a significant difference between the solution of this simplified and result-limited problem and the solution of the real problem. Using intuitive techniques, this drawback of mathematical models is removed. Optimization algorithms have proven to be highly effective in numerous fields, including finance, engineering, and medical. They have been utilized in parameter updates for algorithms like data mining's clustering and classification in addition to these other applications. In several domains, clustering is a commonly utilized technique. Finding the optimal cluster centers for the clustered data is the most crucial step in the clustering process. Intuitive methods provide higher quality solutions by searching in detail instead of limiting the solution space [8]. Intuitive methods do not always guarantee the best case. However, they are called the approximate calculation approach and the aim here is to find acceptable appropriate solutions.

In this study, the Artificial Algae Algorithm

(AAA) has been used to solve clustering problems. In this method, data samples in un-clustered datasets are clustered based on cluster centers obtained through global search. In global search, various probabilities in the search space are examined to minimize the error between data sample clusters. The AAA algorithm has been applied to datasets from the UCI repository, including aggregation, banknote, blobs, ecoli, glass, iris, iris2d, ionosphere, seeds, vertebral2, vertebral3 and wine. The clustering performance of the AAA algorithm compared to the ALO, DEA, MFO, PSO, TSA and WOA algorithms is investigated.

Since the importance of the clustering, there are numerous studies in literature that handled this problem. In their study, Karami and Zapata utilized Particle Swarm Optimization (PSO) and K-Means algorithms to counter attacks in content-based networks. In this approach, input attacks are classified into clusters, and preventive actions are taken based on the results obtained. The authors noted that using the K-Means algorithm alone does not sufficiently optimize cluster centers, so they employed a two-stage approach. In the first stage, called "training stage," cluster centers are obtained using a combination of PSO and K-Means algorithms. In the second stage, called "detection stage," a new fuzzy method is used to identify anomalies in the data. The results show that this method outperforms other clustering algorithms in achieving optimal cluster centers and higher-quality detection rates [9]. In contrast to conventional approaches, Liu and Ban have introduced a novel way to solve the clustering problem that does not need predetermining the number of clusters. The foundation of this approach is the creation of a dynamic, two-dimensional topological graph that shows the connections between the data points in every group. The efficacy and success of the Liu and Ban approach are demonstrated by the experimental results [10]. Rahman and Islam have presented a novel approach to data clustering that combines the K-Means technique with the GA. Finding the ideal number of clusters in the clustering problem is the goal of this approach. The outcomes demonstrate how well our method finds superior cluster centers. The authors showed that, overall, their suggested strategy performs better than five other compared methods after testing it on 20 distinct datasets [11].

Tzortzis and Likas introduced the MinMax K-Means algorithm to overcome the local optimization problem in the classic K-Means algorithm. In K-Means, due to the random selection of initial centroids, the algorithm may get stuck in a local optimum and fail to reach the best possible solution. The MinMax K-Means algorithm addresses this issue by assigning weights to clusters based on variance values, and these weight values are optimized according to the desired objective. Experimental results show that the MinMax K-Means algorithm outperforms the classic K-Means algorithm in terms of accuracy and efficiency [12]. In their study, Maulik and Bandyopadhyay employed GA for data clustering. They used two separate datasets—one artificial and the other real—to test the GA-based clustering technique. The findings show that, on average, the GA-based clustering approach outperformed the K-Means algorithm on the datasets [13]. Van der Merwe and Engelbrecht created a new clustering technique using the PSO algorithm. They compared the PSO-based clustering method's performance with the K-Means approach after evaluating it on six distinct datasets. The outcomes show that in terms of accuracy and efficiency, the PSO-based clustering method performs better than the K-Means approach [14]. The K-Means Algorithm was presented by Shelokar and associates in 2004 as a solution to the clustering problem. Tests have been conducted on both synthetic and actual datasets using this approach. Furthermore, the K-Means algorithm's performance has been contrasted with well-known optimization techniques like GA, Tabu Search, and Simulated Annealing. The outcomes of the experiments indicate that the KA algorithm performs extremely well in data clustering [15]. The PSO technique was used by Omran and his colleagues to update the cluster centers in the K-Means algorithm. Test datasets for image segmentation have been used to evaluate this PSO-based K-Means technique. Furthermore, the PSO-based K-Means algorithm's performance has been contrasted with that of the PSO and K-Means algorithms alone. The PSO-based K-Means algorithm is quite effective at segmenting images, according to experimental findings [16]. Zhang et al. solved the clustering problem by applying the Bee Colony Optimization (BCO) algorithm, which was inspired by the behavior of honeybees. They contrasted the widely used

optimization techniques GA, Simulated Annealing, Tabu Search, and PSO with the BCO-based clustering approach. Three datasets—the iris, wine, and thyroid—that are frequently utilized for clustering were used in this comparison from the UCI repository. The experimental findings demonstrate that, across all datasets, the BCO-based clustering algorithm performs better than alternative optimization-based clustering methods [17]. Mat and associates created an innovative and effective clustering method that imitates the hunting behavior of whales. Ten medical datasets from the UCI Machine Learning repository were clustered using the newly developed WOA-LF technique. The original WOA clustering method, fuzzy c-means, k-means, and k-medoids were compared with WOA-LF's clustering performance. According to the application results, WOA-LF performs better in clustering tasks overall and may be utilized as a substitute method [18].

## 2. Problem Definition

The technique of separating data into distinct groups (clusters) according to their commonalities is known as clustering. These collections of data are referred to as "clusters," and each cluster's data is more comparable to its own than it is to that of other clusters. One popular data mining method for gleaning useful information from massive volumes of data is clustering. Numerous industries, including marketing, bioinformatics, health, and pattern recognition, use clustering. One effective method for identifying structure in data is clustering. Clustering is a helpful method for data analysis in many domains, despite its limitations [19]. The clustering approaches can be categorized into two main topics.

### 2.1. Hierarchical Clustering

Hierarchical clustering methods are used to sequentially determine clusters by bringing units together at different stages and to determine which distance (or similarity) level determines which elements will be members of the clusters. Hierarchical clustering can be examined in two groups: agglomerative hierarchical clustering and divisive hierarchical clustering. Agglomerative hierarchical clustering considers each observation in the data as a cluster. The merging operations are continued until a single cluster is obtained. Divisive hierarchical clustering assumes that all units form a

cluster at the beginning and gradually separates the units into clusters. In hierarchical clustering techniques, clusters are merged sequentially and once a group is merged with another, it is not separated again in subsequent steps. These techniques create a hierarchical structure for the variables under consideration. The number of clusters in hierarchical clustering techniques is decided visually [20].

### 2.2. Non-Hierarchical Clustering

It can be applied if the researcher has determined the number of clusters that will be significant or if the number of clusters is known in advance. The division of units into clusters in this clustering technique can be done at random. The units are assigned to their respective clusters using the clustering criterion once the number of clusters into which they can be divided has been determined. K-means clustering, k-medoids clustering and fuzzy c-means clustering are some examples of non-hierarchical clustering techniques [2], [20].

In this study, artificial algae algorithm and some other metaheuristic algorithms were applied for non-hierarchical clustering.

## 3. Artificial Algae Algorithm (AAA)

AAA is a nature-inspired optimization method that draws inspiration from artificial algal behavior. An artificial alga performs photosynthesis by moving in a helical pattern toward a light source, just like a genuine algae does. It has the ability to shift the dominant species, adapt to its surroundings, and procreate through mitosis. An artificial algal colony represents every solution in the issue space. Each algae colony has an equal amount of algae cells as the problem dimension. The optimum is reached when an algal colony finds the perfect solution. Three primary components comprise the artificial algae algorithm: helical movement, adaptability, and evolutionary process [21].

### 3.1. Helical Movement

Artificial algae cells move helically towards the light. The energy of each helical movement determines whether the colony will change its position in space. At the beginning of each cycle, energy is calculated in proportion to the colony size and this energy represents the quality of the solution. The movement of algae in one dimension is shown in Equation 1 and the movements in the other two dimensions are shown in Equations 2 and 3 [22].

$$x_{im}^{t+1} = x_{im}^t + (x_{jm}^t - x_{im}^t)(sf - \omega_i) \quad (1)$$

$$x_{ik}^{t+1} = x_{ik}^t + (x_{jk}^t - x_{ik}^t)(sf - \omega_i) \cos \alpha \quad (2)$$

$$x_{il}^{t+1} = x_{il}^t + (x_{jl}^t - x_{il}^t)(sf - \omega_i) \sin \beta \quad (3)$$

Here,  $m$ ,  $k$ , and  $l$  are random numbers selected from the range  $[1, d]$ .  $x_i$ ,  $y_i$ , and  $z_i$  represent the  $x$ ,  $y$ , and  $z$  coordinates of the  $i$ th algae colony, respectively.  $j$  is the index of a neighboring algae colony obtained by tournament selection.  $p$  is a real number selected from the range  $[-1, 1]$ .  $\alpha$  and  $\beta$  are randomly selected angles from the range  $[0, 2\pi]$ .  $sf$  is the shear force caused by viscous movement.  $A_i$  is the frictional surface area of the  $i$ th algae colony, which is proportional to its size. The frictional surface is calculated as the surface area of the hemisphere surrounding the algae colony due to its spherical shape. The frictional surface is given by Equation 4.

$$\omega_i = 2\pi r_i^2 \quad (4)$$

$$r_i = \sqrt[3]{\frac{3S_i}{4\pi}} \quad (5)$$

where  $r_i$  is the radius of the hemisphere of the  $i$ th algae colony and  $S_i$  is its volume.

### 3.2. Adaptation

Adaptation is the process by which an algae colony that is not growing sufficiently tries to resemble the largest algae colony in the vicinity. The hunger level determined by the helical movement is used. The hunger level does not change in the colony that goes to a better solution, while the hunger level of the colony that worsens increases. After each helical movement, the colony with the highest hunger value undergoes adaptation (Equation 6 and 7). Whether or not adaptation occurs is determined by the Adaptation parameter (Ap). Ap is a fixed value in the range  $[0, 1]$  and is compared to a random number in this range. If the number is less than the Ap parameter, the adaptation process is carried out [22].

$$starving^t = \operatorname{argmax}\{starvation(x_i)\} \quad (6)$$

$$starving^{t+1} = starving^t + (biggest^t - starving^t) \times rand \quad (7)$$

Here,  $starvation(x_i)$  represents the hunger level

of the  $i$ th algae colony,  $starvingt$  is the algae colony with the highest hunger value at time  $t$ ,  $biggestt$  is the largest algae colony at time  $t$ , and  $rand$  is a real number randomly generated from the range  $[0,1]$ .

### 3.3. Evolutionary Process

Artificial algae cells grow, develop, and divide into two artificial algae cells when they receive sufficient light. Algae cells that do not receive enough light die after a while. The evolutionary process is the stage where an algae cell from the largest algae colony (which has found solutions with better fitness function values than other colonies) is copied to replace each dead cell of the smallest algae colony (which has found solutions with worse fitness function values than other colonies) during the search process. This process is carried out as in Equations 8-10.

$$biggest^t = \arg \max size(x_i^t), i = 1, 2, \dots, N \quad (8)$$

$$smallest^t = \arg \max size(x_i^t), i = 1, 2, \dots, N \quad (9)$$

$$smallest_m^{t+1} = biggest_m^{t+1}, m = 1, 2, \dots, D \quad (10)$$

Here,  $smallest$  represents the smallest algae colony,  $biggest$  represents the largest algae colony, and  $D$  represents the problem dimension.

## 4. Experimental Environment

In this section, the datasets and the comparison metrics are detailed presented.

### 4.1. Datasets

In this study, the performance of AAA was evaluated on biomedical datasets obtained from UCI. The characteristics of the aggregation, banknote, blobs, ecoli, glass, iris, iris2d, ionosphere, seeds, vertebral2, vertebral3 and wine datasets selected from the UCI datasets are given in Table 1.

**Table 1** Datasets used in the study

Dataset	Samples	Attributes	Classes
Aggregation	788	2	7
Banknote	1372	4	2
Blobs	1500	4	4
Ecoli	336	7	8
Glass	214	9	6
Iris	150	4	3
Iris2D	150	2	3
Ionosphere	200	11	6

Seeds	210	7	7
Vertebral2	310	6	2
Vertebral3	310	18	2
Wine	178	13	3

A larger view may be obtained by merging many data pieces, a process known as **aggregation**. Many industries, including research, marketing, finance, and healthcare, can benefit from the usage of aggregation datasets. The Aggregation dataset has 788 data instances overall and includes 2 numerical characteristics and 7 classifications.

The **Banknote** dataset contains images of real banknotes from various currencies. Such datasets are commonly used in the fields of artificial intelligence and image processing for tasks like counterfeit banknote detection, banknote classification, and money counting. This dataset has 4 numerical features and 2 classes, and the Banknote dataset consists of a total of 1372 data instances.

The **Blobs** dataset is a type of dataset that represents clusters of multiple points in a two-dimensional space. These points are often referred to as "blobs" and typically have a circular or elliptical shape. Blobs datasets are commonly used in fields like image processing, computer vision, and machine learning. The Blobs dataset has 4 numerical features and 4 classes, and the Blobs dataset consists of a total of 1500 data instances.

The **Ecoli** dataset is a type of dataset used by researchers who study the characteristics and behavior of the Escherichia coli bacterium. These datasets are commonly used in fields like microbiology, biology, computer science, bioinformatics, and medicine. The E. coli dataset has 7 numerical features and 8 classes, and the E. coli dataset consists of a total of 336 data instances.

The **Glass** dataset is a dataset used to predict glass quality. This dataset contains data about the chemical composition and properties of glass relevant to the glass industry. It is commonly used to train and test machine learning algorithms. The Glass dataset has 9 numerical features and 6 classes, and the Glass dataset consists of a total of 214 data instances.

The **Iris** dataset is a dataset published in 1936 by British statistician and biologist Ronald Fisher, and is considered a classic example of multivariate data analysis. This dataset is commonly used as a test dataset to evaluate and compare the performance of classification algorithms. In particular, it is an ideal

dataset for training and testing supervised learning algorithms because it has labeled data (each instance indicates which flower species it belongs to). This dataset has 4 features and 3 class information, and the values of the features are taken from the width and length of the petals of the iris flower. This dataset consists of 150 examples belonging to three different species of Iris flowers. These examples are equally divided into 3 classes. There are 50 examples in the first class, Iris Setosa, 50 examples in the second class, Iris Versicolour, and finally 50 examples in the Iris Virginica class. The classification of the iris flower is done with these data examples.

The **Iris2D** dataset is a derivative of the Iris dataset and is commonly used for visualization and training of machine learning algorithms. This dataset contains the first two features of the "Iris" dataset (sepal length and sepal width) and is used to understand the performance of classification or clustering algorithms on these features. The "Iris2d" dataset contains the two-dimensional features of each flower instance, such as sepal length and sepal width. This makes it easy to visualize the data on a two-dimensional plane and see how the flower species are grouped based on these two features. This dataset is particularly common for data visualization and evaluating the performance of classification algorithms. For example, it can be used to visualize the outputs of different classification or clustering algorithms to see if the data points are correctly grouped or classified. The Iris2d dataset has 2 numerical features and 3 classes, and the Iris2d dataset consists of a total of 150 data instances.

The **Ionosphere** dataset is particularly used to evaluate the performance of classification algorithms. For example, a machine learning model can try to predict the state of degradation of an ionospheric radar signal using the features in the dataset. The Ionosphere dataset is widely used for training, research, and testing purposes in the fields of machine learning and data mining. The Ionosphere dataset has 11 numerical features and 6 classes, and the Ionosphere dataset consists of a total of 200 data instances.

The **Seeds** dataset is a dataset found in the UCI Machine Learning Repository. This dataset was obtained for use in agricultural research and contains seeds of three different wheat types (Kama, Rosa, and Canadian) in total. This dataset can be used to

evaluate the performance of classification algorithms, to distinguish between wheat types, or for use in agricultural research. The Seeds dataset is particularly commonly used for training and evaluating machine learning classification algorithms. The Seeds dataset has 7 numerical features and 7 classes, and the Seeds dataset consists of a total of 210 data instances.

The **Vertebral2** dataset is a dataset containing information about vertebrae. This dataset can be used to develop artificial intelligence and machine learning models that can be used in the diagnosis and treatment of spinal diseases. The Vertebral2 dataset has 6 numerical features and 2 classes, and the Vertebral2 dataset consists of a total of 310 data instances.

The **Vertebral3** dataset is an important tool in spinal disease research. It can be used to improve the diagnosis, treatment, and prevention of spinal diseases. The Vertebral3 dataset has 18 numerical features and 2 classes, and the Vertebral3 dataset consists of a total of 310 data instances.

The **Wine** dataset is a popular dataset in the fields of machine learning and data science. It can be used for various tasks such as classifying wine types, predicting wine quality, and predicting wine price. It is particularly commonly used in classification and regression problems. The Wine dataset has 13 numerical features and 3 classes, and the Wine dataset consists of a total of 178 data instances.

## 4.2. Comparison Metrics

In this study, three metrics used to compare the performances of the algorithms. On the other hand, sum squared error value was used as fitness function.

### 4.2.1. Sum Squared Error (SSE)

SSE (Sum Squared Error) is a metric used in fields such as statistics and machine learning. It is commonly used to evaluate the performance of regression models. SSE represents the sum of the squares of the differences between the actual values and the predicted values of a model. The primary goal of a regression model is to predict the true values of the dependent variable as accurately as possible. The differences between the predictions and the true values represent the errors. SSE takes the squares of these errors and calculates their sum. Mathematically, for  $n$  data points, SSE is calculated using the following formula:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (11)$$

Here:

- $y_i$  is the true value of the  $i$ th data point,
- $\hat{y}_i$  is the predicted value of the  $i$ th data point by the model,
- $n$  is the total number of data points.

A low SSE value indicates that the model fits the data better and makes better predictions. As the value of SSE approaches zero, the model's predictions get closer to the true values.

#### 4.2.2. Sensitivity and Specificity

The terms "sensitivity" and "specificity" are used in statistics and medicine to quantitatively characterize how well a test detects the existence or absence of a medical condition. If those with the ailment are viewed as "positive" and people without it as "negative," then a test's sensitivity and specificity may be used to determine whether or not it can accurately detect real positives and true negatives, respectively:

- Sensitivity, often known as the true positive rate, is the likelihood of a positive test result provided the subject is in fact positive.

- Specificity, also known as the true negative rate, is the likelihood of a negative test result, provided that the subject is indeed negative.

Sensitivity and specificity can be specified in reference to an assumed-to-be-correct "gold standard test" if the real state of the ailment is unknown. Sensitivity and specificity are typically traded off in testing, both for diagnosis and screening, such that higher sensitivities imply lower specificities and vice versa. A test will be considered highly sensitive if it can consistently identify the existence of a disease, producing a high proportion of genuine positive results and a low percentage of false negative results.

This is crucial when the ailment has major consequences if left untreated and/or when the treatment is extremely successful with few adverse effects. A high specificity test is one that consistently identifies those who do not have the ailment, producing a high percentage of genuine negative results and a low percentage of false positive results. This is particularly crucial in situations when individuals with a diagnosis may be more likely to

undergo tests, incur more costs, experience stigma, worry, etc. Yerushalmy J, introduced the words "sensitivity" and "specificity" to the American biostatistician community [23]. Different definitions exist for laboratory quality control. For instance, according to Saah AJ, Hoover DR, "analytical sensitivity" is the smallest amount of substance in a sample that can be accurately measured by an assay (synonymously to detection limit). "Analytical specificity" is the ability of an assay to measure one specific organism or substance instead of others. However, this essay focuses on the previously mentioned diagnostic sensitivity and specificity [24]. Application for the investigation of screening Consider a research that assesses a test used to check individuals for illnesses. It is either the case that every test taker has the illness or does not. A positive test result would indicate that the subject has the illness, whereas a negative result would indicate that the subject does not. It is possible that the test results do not accurately reflect each subject's current situation. Under those circumstances:

*True positive:* Ill individuals properly classified as ill

*False positive:* When healthy individuals are mistakenly classified as ill

*False negative:* Sick persons were mistakenly recognized as healthy.

*True negative:* Healthy people were accurately identified as healthy.

The test's sensitivity and specificity may be computed once the figures for true positives, false positives, true negatives, and false negatives have been obtained. Any individual with the illness is likely to be identified by the test as positive if it turns out that the sensitivity is high. Conversely, if the test has a high specificity, it is likely to classify as negative any individual who does not have the condition. The methodology for calculating these ratios is discussed on an NIH website [25].

#### *Sensitivity*

Think of a medical test used to diagnose a disease as an example. The capacity of a test to accurately identify sick people among those who actually have the ailment is known as sensitivity, which is also frequently referred to as the detection rate in a clinical environment [26]. Mathematically, this can be expressed as:

$$\text{Sensitivity} = \frac{TP}{P} = \frac{TP}{TP + FN} \quad (12)$$

Since a high sensitivity test seldom misdiagnoses patients who actually have the disease, a negative result can be helpful in "ruling out" sickness [26]. All individuals with the condition will be identified by a test that tests positive and has 100% sensitivity. In this instance, a negative test result would categorically rule out the patient's illness. A high sensitivity test result, however, is not always helpful in "ruling in" a condition. Let's say a "phony" test kit is made to consistently provide a positive result. Test sensitivity is 100% when performed on sick individuals, since all of them test positive. False positives are not considered by sensitivity, though. In addition, the fraudulent test has a false positive rate of 100% on all healthy people, meaning that it is ineffective for "ruling in" or identifying the illness. For the purpose of determining sensitivity, indeterminate test results are ignored. Samples that are uncertain can be treated as false negatives, which yield the worst-case sensitivity value and may cause it to be underestimated, or they can be eliminated from the analysis (it is important to specify the number of exclusions when mentioning sensitivity).

### *Specificity*

Examine a medical test as an illustration to help you understand the concept. The capacity of the test to accurately rule out healthy individuals in the absence of a problem is referred to as specificity. Here's how to write it mathematically:

$$\text{Specificity} = \frac{TN}{N} = \frac{TN}{TN + FP} \quad (13)$$

Since a test with high specificity seldom yields positive findings in healthy individuals, a positive result might be helpful for "ruling in" illness [27]. A positive test result would categorically rule in the presence of the disease since a test with 100% specificity would identify all patients who do not have the condition by testing negative. However, "ruling out" a condition is not always possible with a negative result from a test with great specificity. Since specificity does not account for erroneous negative results, a test that consistently yields a negative result, for instance, would have a specificity of 100%. Such a test would not be useful for "ruling

out" the condition since it would yield a negative result for those who already had the illness.

### **4.2.3. F1-Score**

The F1-score is a performance measure commonly used in classification problems. This metric is used to evaluate the accuracy of a model, particularly useful in imbalanced classification problems. The F1-score is a combined value of the precision and recall metrics. The percentage of positive examples among those that a model predicts to be positive is known as precision. The percentage of genuinely positive cases that the model accurately predicts as positive is known as recall. The F1-score is calculated using the following formula, using precision (P) and recall (R) values:

$$F = 2 \cdot \frac{P \cdot R}{P + R} \quad (14)$$

The F1-score represents a balance between precision and recall. It is particularly used in classification problems where there is an imbalance between classes. The F1-score takes a value between 0 and 1, with a value closer to 1 representing better performance.

## **5. Results**

In this study, the Artificial Algae Algorithm was used to classify 12 datasets (aggregation, banknote, blobs, ecoli, glass, iris, iris2d, ionosphere, seeds, vertebral2, vertebral3 and wine) obtained from the UCI repository. The aim of the algorithms was to identify the clusters that minimize the values of the SSE. The algorithms AAA, ALO, DEA, MFO, PSO, TSA and WOA were run for each dataset with a population size of 40, a maxFEs (maximum fitness evaluation size) of 10000 and a runtime of 30. The values of the F1-Score, Sensitivity and Specificity metrics obtained as a result of this study are shown in the tables below. The three tables below show the F1-Score, Sensitivity and Specificity values obtained in 30 different runs of seven different algorithms for each dataset, as well as the average (Mean) and standard deviation (Std.) of these values. In addition, the "Rank" column is used to indicate the performance of each algorithm. F1-Score, Sensitivity and Specificity are metrics that measure the accuracy of a clustering algorithm, with higher values



indicating better performance. *Note: When writing the results, 2 digits after the comma were taken as precision. Therefore, although the algorithms appear to have the same metric value, their rank may appear different.*

Table 2 shows the results of the algorithms for F1-score metric. According to the Table 2, the AAA algorithm has an average F1-score value of 0.72 for 12 datasets and ranks first in the rank column. The ALO algorithm ranks first in the rank column for the blobs, iris, and iris2d datasets. The MFO and TSA algorithms rank first in the rank column for the banknote, blobs, and iris2d datasets, and the WOA algorithm ranks first in the rank column for the iris2d

dataset. In general, Table 2 shows that the AAA algorithm ranks first with an average success ranking of 1, the MFO algorithm ranks second with an average ranking of 2.5, the ALO algorithm ranks third with an average ranking of 2.66, the TSA algorithm ranks fourth with an average ranking of 3.08, the WOA algorithm ranks fifth with an average ranking of 4.16, the DEA algorithm ranks sixth with an average ranking of 4.91, and the PSO algorithm ranks seventh with an average ranking of 5.5.

**Table 2.** Experimental results of AAA and other algorithms for F1-Score metric

Datasets		AAA	ALO	DEA	MFO	PSO	TSA	WOA
Aggregation	Mean	0.81	0.73	0.71	0.77	0.62	0.71	0.67
	Std	0.04	0.10	0.03	0.07	0.14	0.09	0.11
	Rank	1	3	4	2	7	5	6
Banknote	Mean	0.79	0.79	0.46	0.79	0.45	0.79	0.75
	Std	0.00	0.01	0.22	0.00	0.16	0.00	0.08
	Rank	1	2	4	1	5	1	3
Blobs	Mean	1.00	1.00	1.00	1.00	0.83	1.00	0.99
	Std	0.00	0.00	0.00	0.00	0.27	0.00	0.08
	Rank	1	1	2	1	4	1	3
Ecoli	Mean	0.39	0.31	0.29	0.35	0.25	0.29	0.30
	Std	0.11	0.07	0.08	0.12	0.09	0.09	0.09
	Rank	1	3	2	2	7	6	4
Glass	Mean	0.24	0.14	0.14	0.20	0.18	0.18	0.13
	Std	0.06	0.05	0.03	0.06	0.06	0.04	0.03
	Rank	1	5	6	2	3	4	7
Iris	Mean	0.96	0.96	0.83	0.89	0.68	0.95	0.84
	Std	0.00	0.00	0.09	0.12	0.21	0.01	0.18
	Rank	1	1	5	3	6	2	4
Iris2d	Mean	0.96	0.96	0.94	0.96	0.88	0.96	0.96
	Std	0.00	0.00	0.03	0.00	0.10	0.00	0.00
	Rank	1	1	2	1	3	1	1
Ionosphere	Mean	0.69	0.67	0.40	0.54	0.54	0.67	0.67
	Std	0.00	0.03	0.04	0.13	0.10	0.02	0.05
	Rank	1	4	7	5	6	3	2
Seeds	Mean	0.87	0.83	0.37	0.55	0.29	0.56	0.31
	Std	0.02	0.16	0.15	0.27	0.14	0.25	0.18
	Rank	1	2	5	4	7	3	6
Vertebral2	Mean	0.65	0.53	0.39	0.57	0.41	0.39	0.44
	Std	0.00	0.14	0.02	0.12	0.05	0.13	0.11
	Rank	1	3	6	2	5	7	4
Vertebral3	Mean	0.38	0.21	0.15	0.23	0.15	0.26	0.18
	Std	0.03	0.08	0.04	0.08	0.05	0.08	0.09
	Rank	1	4	6	3	7	2	5
Wine	Mean	0.90	0.39	0.24	0.37	0.31	0.46	0.34
	Std	0.03	0.20	0.15	0.21	0.19	0.14	0.22
	Rank	1	3	7	4	6	2	5
Average Rank		1	2.67	4.92	2.5	5.5	3.08	4.17

Table 3 shows the results of the algorithms for Sensitivity metric. According to the Table 3, the AAA algorithm has an average Sensitivity value of

0.75 for 12 datasets and ranks first in the rank column. The ALO algorithm ranks first in the rank column for the blobs, iris, and iris2d datasets. The

MFO and TSA algorithms rank first in the rank column for the banknote, blobs, and iris2d datasets, and the WOA algorithm ranks first in the rank column for the iris2d dataset. In general, Table 3 shows that the AAA algorithm ranks first with an average success ranking of 1, the MFO algorithm ranks second with an average ranking of 2.5, the TSA

algorithm ranks third with an average ranking of 2.83, the ALO algorithm ranks fourth with an average ranking of 2.91, the WOA algorithm ranks fifth with an average ranking of 4.25, the DEA algorithm ranks sixth with an average ranking of 5.08, and the PSO algorithm ranks seventh with an average ranking of 5.25.

**Table 3.** Experimental results of AAA and other algorithms for Sensitivity metric

Datasets		AAA	ALO	DEA	MFO	PSO	TSA	WOA
Aggregation	Mean	0.90	0.80	0.77	0.86	0.67	0.78	0.74
	Std	0.06	0.12	0.06	0.09	0.16	0.11	0.12
	Rank	1	3	5	2	7	4	6
Banknote	Mean	0.80	0.80	0.50	0.80	0.53	0.80	0.76
	Std	0.00	0.01	0.20	0.00	0.12	0.00	0.08
	Rank	1	2	5	1	4	1	3
Blobs	Mean	1.00	1.00	1.00	1.00	0.85	1.00	0.99
	Std	0.00	0.00	0.00	0.00	0.24	0.00	0.06
	Rank	1	1	2	1	4	1	3
Ecoli	Mean	0.41	0.37	0.32	0.39	0.28	0.33	0.34
	Std	0.12	0.09	0.09	0.13	0.10	0.10	0.12
	Rank	1	3	6	2	7	5	4
Glass	Mean	0.30	0.22	0.20	0.28	0.24	0.24	0.20
	Std	0.05	0.05	0.02	0.07	0.06	0.04	0.03
	Rank	1	5	7	2	3	4	6
Iris	Mean	0.96	0.96	0.84	0.90	0.72	0.95	0.87
	Std	0.00	0.00	0.07	0.10	0.17	0.01	0.14
	Rank	1	1	5	3	6	2	4
Iris2d	Mean	0.96	0.96	0.94	0.96	0.88	0.96	0.96
	Std	0.00	0.00	0.03	0.00	0.09	0.00	0.00
	Rank	1	1	2	1	3	1	1
Ionosphere	Mean	0.71	0.67	0.50	0.58	0.56	0.68	0.68
	Std	0.00	0.03	0.03	0.11	0.08	0.02	0.04
	Rank	1	4	7	5	6	2	3
Seeds	Mean	0.87	0.84	0.46	0.59	0.40	0.60	0.39
	Std	0.02	0.14	0.12	0.23	0.12	0.22	0.16
	Rank	1	2	5	4	6	3	7
Vertebral2	Mean	0.72	0.59	0.47	0.64	0.48	0.40	0.51
	Std	0.00	0.14	0.05	0.12	0.05	0.16	0.12
	Rank	1	3	6	2	5	7	4
Vertebral3	Mean	0.45	0.34	0.35	0.36	0.35	0.38	0.35
	Std	0.02	0.05	0.03	0.05	0.02	0.05	0.05
	Rank	1	7	4	3	6	2	5
Wine	Mean	0.91	0.54	0.40	0.51	0.44	0.57	0.47
	Std	0.02	0.18	0.13	0.19	0.15	0.13	0.18
	Rank	1	3	7	4	6	2	5
Average Rank		<b>1</b>	2.17	5.08	2.5	5.25	2.83	4.25

Table 4 shows the results of the algorithms for Specificity metric. According to the Table 4, the AAA algorithm ranks first in the rank column for 11 datasets and ranks second in the rank column with an average Specificity value of 0.94 for the ecoli dataset. The ALO algorithm ranks first in the rank column for the blobs, ecoli, iris, and iris2d datasets. The MFO and TSA algorithms rank first in the rank column for

the banknote, blobs, and iris2d datasets, and the WOA algorithm ranks first in the rank column for the iris2d dataset. In general, Table 4 shows that the AAA algorithm ranks first with an average success ranking of 1.08, the ALO and MFO algorithms rank second with an average ranking of 2.75, the TSA algorithm ranks third with an average ranking of 2.83, the WOA algorithm ranks fourth with an

average ranking of 4.41, the DEA algorithm ranks fifth with an average ranking of 4.91, and the PSO algorithm ranks sixth with an average ranking of 5.08.

**Table 4.** Experimental results of AAA and other algorithms for Specificity metric

Datasets		AAA	ALO	DEA	MFO	PSO	TSA	WOA
Aggregation	Mean	0.98	0.97	0.97	0.97	0.96	0.96	0.96
	Std	0.00	0.01	0.01	0.01	0.02	0.01	0.02
	Rank	1	3	4	2	7	5	6
Banknote	Mean	0.80	0.80	0.50	0.80	0.53	0.80	0.76
	Std	0.00	0.01	0.20	0.00	0.12	0.00	0.08
	Rank	1	2	5	1	4	1	3
Blobs	Mean	1.00	1.00	1.00	1.00	0.92	1.00	0.99
	Std	0.00	0.00	0.00	0.00	0.12	0.00	0.03
	Rank	1	1	2	1	4	1	3
Ecoli	Mean	0.94	0.94	0.92	0.94	0.92	0.93	0.93
	Std	0.01	0.01	0.02	0.02	0.02	0.01	0.02
	Rank	2	1	6	3	7	5	4
Glass	Mean	0.86	0.85	0.84	0.86	0.85	0.85	0.84
	Std	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	Rank	1	5	6	2	4	3	7
Iris	Mean	0.98	0.98	0.92	0.95	0.86	0.98	0.93
	Std	0.00	0.00	0.04	0.05	0.09	0.01	0.07
	Rank	1	1	5	3	6	2	4
Iris2d	Mean	0.98	0.98	0.97	0.98	0.94	0.98	0.98
	Std	0.00	0.00	0.01	0.00	0.05	0.00	0.00
	Rank	1	1	2	1	3	1	1
Ionosphere	Mean	0.71	0.67	0.50	0.58	0.56	0.68	0.68
	Std	0.00	0.03	0.03	0.11	0.08	0.02	0.04
	Rank	1	4	7	5	6	2	3
Seeds	Mean	0.94	0.92	0.73	0.80	0.70	0.80	0.70
	Std	0.01	0.07	0.06	0.12	0.06	0.11	0.08
	Rank	1	2	5	4	6	3	7
Vertebral2	Mean	0.72	0.59	0.47	0.64	0.48	0.40	0.51
	Std	0.00	0.14	0.05	0.12	0.05	0.16	0.12
	Rank	1	3	6	2	5	7	4
Vertebral3	Mean	0.72	0.66	0.67	0.67	0.67	0.68	0.67
	Std	0.01	0.03	0.02	0.03	0.02	0.03	0.03
	Rank	1	7	4	5	3	2	6
Wine	Mean	0.95	0.75	0.70	0.74	0.71	0.76	0.73
	Std	0.01	0.08	0.06	0.08	0.07	0.05	0.08
	Rank	1	3	7	4	6	2	5
Average Rank		<b>1.08</b>	2.75	4.92	2.75	5.08	2.83	4.42

When the results are examined in general, it is seen that AAA outperforms other algorithms. This result shows that the local and global search strategies of the algorithm are suitable for the problems studied. In addition, in the position update process of the algorithm, the elimination of bad members and the inclusion of members with better fitness values into the population positively affects its success.

## 6. Discussion and Conclusion

In this study, the accuracy of clustering was improved by using AAA to obtain appropriate cluster centers and increase clustering success. The

performance of AAA was evaluated using 12 commonly used datasets (aggregation, banknote, blobs, ecoli, glass, iris, iris2d, ionosphere, seeds, vertebral2, vertebral3 and wine) from the UCI repository. AAA's performance was compared with the performance of the six metaheuristic algorithms by using three metrics: F1-Score, Sensitivity and Specificity. The experimental results obtained show that AAA is more successful than the other algorithms in the clustering problem.

For future studies, the performance of the algorithms can be increased by using hybrid metaheuristic algorithms. In addition, the

performance of the algorithms can be evaluated by using data sets with different characteristics.

## References

- [1] K. Özdamar, *Paket Programlar İle İstatistiksel Veri Analizi 1*. Kaan Kitabevi, 1997.
- [2] H. Tatlıdil, *Uygulamalı Çok Değişkenli İstatistiksel Analiz*. Accessed: Oct. 02, 2024. [Online]. Available: <https://www.nadirkitap.com/uygulamali-cok-degiskenli-istatistiksel-analiz-prof-dr-huseyin-tatlidil-kitap1444523.html>
- [3] J. F. Hair Jr., W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate Data Analysis*. Accessed: Oct. 02, 2024. [Online]. Available: <https://www.scirp.org/reference/ReferencesPapers?ReferenceID=1519308>
- [4] M. Lorr, *Cluster Analysis for Social Scientists*, 1st edition. San Francisco: Jossey-Bass Inc Pub, 1983.
- [5] S. Sharma, *Applied Multivariate Techniques*, 1st edition. New York: Wiley, 1995.
- [6] J. Tabak, *Geometry: the language of space and form*, Rev. ed. in *The history of mathematics*. New York, NY: Facts On File, 2011.
- [7] S. Ishak Boushaki, N. Kamel, and O. Bendjeghaba, 'A new quantum chaotic cuckoo search algorithm for data clustering', *Expert Systems with Applications*, vol. 96, Dec. 2017, doi: 10.1016/j.eswa.2017.12.001.
- [8] X.-S. Yang, 'A New Metaheuristic Bat-Inspired Algorithm', vol. 284, Apr. 2010, doi: 10.1007/978-3-642-12538-6\_6.
- [9] A. Karami and M. Guerrero-Zapata, 'A fuzzy anomaly detection system based on hybrid PSO-Kmeans algorithm in content-centric networks', *Neurocomputing*, vol. 149, pp. 1253–1269, Feb. 2015, doi: 10.1016/j.neucom.2014.08.070.
- [10] H. Liu and X. Ban, 'Clustering by growing incremental self-organizing neural network', *Expert Systems with Applications*, vol. 42, no. 11, pp. 4965–4981, Jul. 2015, doi: 10.1016/j.eswa.2015.02.006.
- [11] M. A. Rahman and M. Z. Islam, 'A hybrid clustering technique combining a novel genetic algorithm with K-Means', *Knowledge-Based Systems*, vol. 71, pp. 345–365, Nov. 2014, doi: 10.1016/j.knsys.2014.08.011.
- [12] G. Tzortzis and A. Likas, 'The MinMax  $k$ -Means clustering algorithm', *Pattern Recognition*, vol. 47, no. 7, pp. 2505–2516, Jul. 2014, doi: 10.1016/j.patcog.2014.01.015.
- [13] U. Maulik and S. Bandyopadhyay, 'Genetic algorithm-based clustering technique', *Pattern Recognition*, vol. 33, no. 9, pp. 1455–1465, Sep. 2000, doi: 10.1016/S0031-3203(99)00137-5.
- [14] D. Merwe and A. Engelbrecht, 'Data clustering using particle swarm optimization[C]', presented at the Proc of 2003 Congress on Evolutionary Computation (CEC'03), Jan. 2003, pp. 215–220. doi: 10.1109/CEC.2003.1299577.
- [15] P. S. Shelokar, V. K. Jayaraman, and B. D. Kulkarni, 'An ant colony approach for clustering', *Analytica Chimica Acta*, vol. 509, no. 2, pp. 187–195, May 2004, doi: 10.1016/j.aca.2003.12.032.
- [16] M. Omran, A. Engelbrecht, and A. Salman, 'Dynamic Clustering using Particle Swarm Optimization with Application in Unsupervised Image Classification', vol. 9, Jan. 2005.
- [17] C. Zhang, D. Ouyang, and J. Ning, 'An artificial bee colony approach for clustering', *Expert Systems with Applications*, vol. 37, no. 7, pp. 4761–4767, Jul. 2010, doi: 10.1016/j.eswa.2009.11.003.
- [18] A. N. Mat, O. İnan, and M. Karakoyun, 'An application of the whale optimization algorithm with Levy flight strategy for clustering of medical datasets', *An International Journal of Optimization and Control: Theories & Applications (IJOCTA)*, vol. 11, no. 2, Art. no. 2, Jun. 2021, doi: 10.11121/ijocta.01.2021.001091.
- [19] G. Sarıman, 'Veri Madenciliğinde Kümeleme Teknikleri Üzerine Bir Çalışma: K-Means ve K-Medoids Kümeleme Algoritmalarının Karşılaştırılması'.
- [20] B. S. Everitt and G. Dunn, *Applied Multivariate Data Analysis*. Oxford University Press, 1992.
- [21] S. A. Uymaz, G. Tezel, and E. Yel, 'Artificial algae algorithm (AAA) for nonlinear global optimization', *Applied Soft Computing*, vol. 31, pp. 153–171, Jun. 2015, doi: 10.1016/j.asoc.2015.03.003.
- [22] X. Zhang et al., 'Binary Artificial Algae Algorithm for Multidimensional Knapsack Problems', *Applied Soft Computing*, vol. 43, Mar. 2016, doi: 10.1016/j.asoc.2016.02.027.
- [23] J. Yerushalmy, 'Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques', *Public Health Rep (1896)*, vol. 62, no. 40, pp. 1432–1449, Oct. 1947.
- [24] A. J. Saah and D. R. Hoover, '[Sensitivity and specificity revisited: significance of the terms in analytic and diagnostic language]', *Ann Dermatol Venereol*, vol. 125, no. 4, pp. 291–294, Apr. 1998.
- [25] R. Parikh, A. Mathai, S. Parikh, G. Chandra Sekhar, and R. Thomas, 'Understanding and using sensitivity, specificity and predictive values', *Indian J Ophthalmol*, vol. 56, no. 1, pp. 45–50, 2008, doi: 10.4103/0301-4738.37595.
- [26] D. G. Altman and J. M. Bland, 'Diagnostic tests. 1: Sensitivity and specificity', *BMJ*, vol. 308, no. 6943, p. 1552, Jun. 1994, doi: 10.1136/bmj.308.6943.1552.
- [27] 'SpPin and SnNout'. Accessed: Oct. 02, 2024. [Online]. Available: <https://www.cebm.ox.ac.uk/resources/ebm-tools/sppin-and-snnout>