Optimizing Clustering Performance: A Novel Integration of Whale Optimization Algorithm and K-NN Validation in Data Mining Analytics

Nur Wahyu Hidayat [a,1,*], Mursalim [b,2], Umar Ghoni [a,3]

[a] Muhammadiyah University of Brebes, Diponegoro Street Nu.184 Grengseng Paguyangan, Brebes and 52276, Indonesia
[b] University of Sugeng Hartono, Ir. Soekarno street Madegondo Grogol, Sukoharjo and 57552, Indonesia
[1] nur.wahyu@umbs.ac.id *; [2] mursalim.dsc@sugenghartono.ac.id;; [3] umar.ghoni@umbs.ac.id
* Author correspondence
Submission:25/04/2025, Revision: 25/04/2025, Accepted: 29/04/2025

Abstract

The digital era's massive data necessitates effective clustering, a machine learning technique grouping data by similarity. Clustering large, complex datasets faces challenges like volume, dimensionality, and variability, hindering algorithms like K-Means. A key issue in K-Means is its sensitivity to initial centroid selection, impacting results. This research aims to optimize clustering performance by integrating the Whale Optimization Algorithm (WOA) for improved initial centroid determination in K-Means, and K-Nearest Neighbors (K-NN) for validating the resulting cluster quality through classification accuracy. Evaluation on iris, wine, heart, lung, and liver datasets using the Davies-Bouldin Index (DBI) showed that WOA-KMeans consistently yielded lower DBI values compared to standard K-Means, indicating superior clustering. Notably, DBI for the lung dataset drastically decreased from 2.38016 to 0.65395. Furthermore, K-NN classification using the generated cluster labels achieved high accuracy (98-99% across datasets), confirming well-separated and internally homogeneous clusters. This demonstrates WOA's effectiveness in guiding K-Means towards better solutions and K-NN's utility in validating cluster distinctiveness. This novel WOA-K-NN combination offers a more accurate and robust clustering method. The significant performance improvements observed across diverse datasets highlight its potential for enhanced data exploration and pattern discovery in complex data mining tasks.
Keywords: whale optimization algorithm, k-means, k-nearest neighbors, davies-bouldin index, accuracy

Abstrak

Data besar era digital memerlukan pengelompokan yang efektif, sebuah teknik pembelajaran mesin yang mengelompokkan data berdasarkan kesamaan. Pengelompokan kumpulan data yang besar dan kompleks menghadapi tantangan seperti volume, dimensionalitas, dan variabilitas, yang menghambat algoritma seperti K-Means. Masalah utama dalam K-Means adalah sensitivitasnya terhadap pemilihan centroid awal, yang memengaruhi hasil. Penelitian ini bertujuan untuk mengoptimalkan kinerja pengelompokan dengan mengintegrasikan Whale Optimization Algorithm (WOA) untuk penentuan centroid awal yang lebih baik dalam K-Means, dan K-Nearest Neighbors (K-NN) untuk memvalidasi kualitas klaster yang dihasilkan melalui akurasi klasifikasi. Evaluasi pada kumpulan data iris, anggur, jantung, paru-paru, dan hati menggunakan Indeks Davies-Bouldin (DBI) menunjukkan bahwa WOA-KMeans secara konsisten menghasilkan nilai DBI yang lebih rendah dibandingkan dengan K-Means standar, yang menunjukkan pengelompokan yang unggul. Khususnya, DBI untuk kumpulan data paru-paru menurun drastis dari 2,38016 menjadi 0,65395. Lebih jauh lagi, klasifikasi K-NN menggunakan label klaster yang dihasilkan mencapai akurasi tinggi (98-99% di seluruh kumpulan data), yang mengonfirmasi klaster yang terpisah dengan baik dan homogen secara internal. Hal ini menunjukkan efektivitas WOA dalam mengarahkan K-Means menuju solusi yang lebih baik dan utilitas K-NN dalam memvalidasi kekhasan klaster. Kombinasi WOA-K-NN yang baru ini menawarkan metode pengelompokan yang lebih akurat dan tangguh. Peningkatan kinerja signifikan yang diamati di seluruh kumpulan data yang beragam menyoroti potensinya untuk eksplorasi data yang lebih baik dan penemuan pola dalam tugas penambangan data yang kompleks..

Kata kunci: algoritma optimisasi ikan paus, k-means, k-nearest neighbors, dbi, akurasi.

1. Introduction

In the current digital era, we are faced with an explosion of data that is so massive [1]. This data contains valuable information that can be utilized for various purposes, one of which is data clustering [2]. Clustering is a technique in machine learning that aims to group data based on the similarity of its characteristics [1]. Imagine we categorize fruits based on their types (apples, oranges, bananas) or customers based on their purchasing preferences [2].

However, performing clustering on a very large and complex dataset is not an easy task [3]. Some challenges that are often faced are:

- Very large data volume: A dataset that is too large can make it difficult for computers to process it in a short time [1].
- High data dimensionality: The more features or variables in the data, the harder it is to find relevant patterns [3].
- High data variability: The presence of noise, outliers, or class imbalance in the data can hinder the performance of clustering algorithms [3].

To address these challenges, various clustering algorithms have been developed, such as K-Means, DBSCAN, and hierarchical clustering[2]. However, each algorithm has its own advantages and disadvantages. One important aspect of clustering is performance optimization, which is how we can find the most optimal and representative data groups[1]. The K-Means algorithm is one of the most popular clustering algorithms that works by iterating data partitions until convergence is reached[3].

Although the K-Means algorithm is the most popular, it has a major weakness: sensitivity to the initial centroid initialization. Different initial centroids can produce very different clustering results [4], [5]. In this research, we are interested in optimizing clustering performance by combining two powerful algorithms, namely the Whale Optimization Algorithm (WOA) and K-Nearest Neighbors (K-NN) [6]. WOA is a metaheuristic algorithm inspired by the behavior of humpback whales in search of prey[7]. This algorithm has excellent capabilities in finding optimal solutions in a vast search space[8]. The WOA algorithm is used in determining the initial centroid in K-Means [4]. Meanwhile, K-NN is a simple yet effective classification algorithm[9]. K-NN can be used to validate clustering results by measuring the classification accuracy of data points based on the obtained cluster labels[10].

This research is expected to contribute to the field of data mining by offering a new approach to clustering optimization. By combining the strengths of WOA in optimization and K-NN in validation, we hope to develop a clustering method that is more accurate, efficient, and robust in handling various types of datasets.

The uniqueness of this research lies in the innovative combination of WOA and K-NN for clustering problems. WOA will be used to optimize clustering parameters, specifically the initialization of centroids from the K-Means algorithm [4], while K-NN will be used to validate the clustering results[11]. This combination is expected to produce better clustering solutions compared to using conventional clustering algorithms.

Whale Optimization Algorithm (WOA)

Whale Optimization Algorithm (WOA) is a metaheuristic algorithm inspired by the hunting behavior of humpback whales [7]. WOA was first introduced by Mirjalili et al. WOA simulates three main stages in the hunting of humpback whales, namely encircling prey, attacking prey, and searching for prey randomly. WOA has demonstrated good performance in various optimization problems, including combinatorial and continuous optimization problems [12]. WOA can also be combined with other metaheuristic algorithm, such as BAT [13].

The mathematical model of WOA [7] can be explained as follows: the algorithm starts with the assumption that the current best solution is either capturing its prey or being close to its prey. All whales update their positions relative to the best whale with the equation: umum.

$$\vec{D} = \left|\vec{C}.\overrightarrow{X^*}(t) - \vec{X}(t)\right| \quad (1)$$

$$\vec{X}(t+1) = \overrightarrow{X^*}(t) - \vec{A}.\vec{D} \quad (2)$$

Where X* is the best whale. X is the current whale, and D is the distance. As for vectors A and C, they are obtained from the following equations:

$$\vec{A} = 2.a.\vec{r} - \vec{a} \quad (3)$$

$$\vec{C} = 2.\vec{r} \quad (4)$$

Where a decreases linearly from 2 to 0, and r is a random vector with values in the range [0, 1]. During exploitation, the whales follow a shrinking mechanism or position update in the form of a spiral, which is referred to as a helix-shaped movement represented by the equation:

$$\vec{X}(t+1) = \overrightarrow{D'}.e^{bl}.\cos(2l) + \overrightarrow{X^*}(t) \quad (5)$$

$$\overrightarrow{D'} = \left|\overrightarrow{X^*}(t) - \vec{X}(t)\right| \quad (6)$$

Where D' is the distance between the prey and the whale, b is the constant for the spiral, and l is a random number in the range [-1, 1]. Searching in a wider area is used in the exploration phase using the equation.

$$\vec{X}(t+1) = \overrightarrow{X_{rand}} - \vec{A}.\vec{D} \quad (7)$$

$$\overrightarrow{D'} = \left|\vec{C}.\overrightarrow{X_{rand}} - \vec{X}\right| \quad (8)$$

The choice between the shrinking mechanism or the spiral equation has an equal chance. The value of p determines the selection of one over the other, as can be seen in the following equation:

$$\bar{X}(t+1) = \begin{cases} \overrightarrow{X^*}(t) - \vec{A}.\vec{D}, if\ p < 0.5 \\ \overrightarrow{D'}.e^{bl}.\cos(2l) + \overrightarrow{X^*}(t), if\ \ p \geq 0.5 \end{cases} \quad (9)$$

Where p is a random value. The two-phase selection determines the balance between intensification and diversification techniques and is applied 50% of the time. In general, the structure and operation of WOA are simple, which facilitates its improvement.

The advantage of WOA in clustering optimization lies in its ability to search for global optimal solutions [12]. By simulating whale hunting behavior, WOA can explore a vast solution space and avoid getting trapped in local minima. In addition, WOA is also relatively easy to implement and has few parameters that need to be adjusted. In the context of clustering, WOA can be used to optimize the initial centroid positions of K-Means, thereby improving the quality of clustering [12]

2. Related Works

In this study, we propose a new approach to optimize clustering performance by integrating the Whale Optimization Algorithm (WOA) and K-Nearest Neighbor (K-NN). WOA will be used to optimize the initial centroid positions of K-Means, while K-NN will be used to validate the clustering results. The quality of clustering will be evaluated using the Davies-Bouldin Index (DBI) and accuracy.

2.1. Whale Optimization Algorithm (WOA) to Optimize Initial Centroid of K-Means

WOA, which mimics whale hunting behavior through the *encircling prey* (the algorithm models the behavior of whales surrounding their prey), *bubble-net attacking* (with *shrinking encircling* and *spiral updating*), and *search for prey* phases, is used to find positions representing the initial K-Means centroids. Each whale's position is a candidate centroid set (two in the code, expandable for *k* clusters).

The quality of each candidate is evaluated using a fitness function that is the negative of the K-Means inertia. The whale population moves within the data feature space based on WOA mechanisms, constrained by parameters such as the number of whales, iterations, and position boundaries. The goal is to find an initial centroid set that minimizes the K-Means inertia, thus resulting in better clustering.

Algorithm: K-Means Centroid Optimization using Whale Optimization Algorithm (WOA)

Input:

- X: Dataset with n samples and d dimensions.
- K: Desired number of clusters (explicitly set to 2 in this code).
- Nwhales: Number of whale population.
- Tmax: Maximum number of iterations for WOA.
- lower_bound: Lower bound for centroid positions.
- upper_bound: Upper bound for centroid positions.

Output:

- Cbest: Best set of centroids found by WOA.

Steps:

1. **Initialize Whale Population:**
    - Generate an initial population of Nwhales whales with random positions within the centroid search space.
    - Each whale's position Pi (for i=1,2,...,Nwhales) represents K initial centroids. In the given code, K=2, so each whale's position has dimensions (2,d), where each row represents a centroid with d dimensions.
    - Ensure each element in the whale's position is within the bounds [lower_bound, upper_bound] for each dimension.
2. **Evaluate Initial Fitness:**
    - For each whale Pi in the population:
        - Run the K-Means algorithm with Pi as the initial centroid initialization (init=population[i], n_init=1).

- Train the K-Means model on the dataset X to obtain cluster labels and inertia (sum of squared distances to the nearest centroid).
- Calculate the fitness value fi of whale Pi as the negative of the inertia (because WOA aims to maximize fitness, and lower inertia indicates better clustering). fi=−inertiai

3. **Find the Best Whale:**
   - Identify the whale with the highest fitness value in the current population. Designate its position as the best position Pbest and its fitness as fbest.
4. **Optimization Iteration:**
   - For each iteration t from 1 to Tmax:
     - Calculate the parameter a: a=2−2×Tmaxt
     - For each whale Pi in the population:
       - Generate random numbers r1 and r2 uniformly distributed in [0,1].
       - Calculate the coefficients A and C: A=2×a×r1−a C=2×r2
       - Generate a random number p uniformly distributed in [0,1].
       - **Exploitation Phase (Encircling Prey and Bubble-net Attacking):**
         - If |A|<1:
           - Calculate the distance vector D between the current whale and the best whale: D=|C×Pbest−Pi|
           - Update the current whale's position: Pit+1=Pbest−A×D
       - **Exploration Phase (Search for Prey):**
         - If |A|≥1:
           - Select a random whale Prand from the current population.
           - Calculate the distance vector Drand between the current whale and the random whale: Drand=|C×Prand−Pi|
           - Update the current whale's position: Pit+1=Prand−A×Drand
       - **Spiral Pattern (Bubble-net Attacking - Alternative):**
         - Generate a random number l uniformly distributed in [−1,1].
         - If p<0.5:
           - Calculate the distance vector Dbest between the current whale and the best whale: Dbest=|Pbest−Pi|
           - Update the current whale's position using the spiral equation: Pit+1=Dbest×ebl×cos (2πl) + Pbest (In the code, b=1 is assumed)
       - **Boundary Handling:**
         - Ensure the new position of the whale Pit+1 remains within the bounds [lower_bound, upper_bound] for each dimension. If any element goes out of bounds, reset it to the nearest bound.
     - **Evaluate Iteration Fitness:**
       - For each whale Pit+1 in the new population:
         - Run K-Means with Pit+1 as the initial centroid.
         - Calculate the fitness value fit+1 (negative of the inertia).
       - **Update Best Whale:**
         - If a new whale has a fitness value higher than fbest, update Pbest with the position of this new whale and fbest with its fitness value.
5. **Stopping Criterion:**
   - Repeat step 4 until the maximum number of iterations Tmax is reached.
6. **Result:**
   - After Tmax iterations, return the best whale position Pbest as the optimized set of centroids.

2.2. Cluster Validation Based on K-Nearest Neighbors (K-NN Validation)

K-Nearest Neighbor (K-NN) is a simple yet effective classification algorithm[14]. K-NN works by classifying a data point based on the labels of its K nearest neighbors. In the context of clustering, K-NN can be used to validate the clustering results[15] in the following way:

1. Labeling Cluster: Each cluster is labeled based on the majority class of the data contained within it.
2. Test Data Classification: Test data is classified using the K-NN algorithm based on the predetermined cluster labels.
3. Accuracy Evaluation: Classification accuracy can be used as a metric to measure the quality of clustering.

To evaluate the quality of the data clustering produced by the combination of the WOA and K-Means algorithms, we employed the K-Nearest Neighbors (K-NN) method. While K-NN is typically used for classifying data, here we utilized it to assess how well this algorithm could separate the already formed data clusters. If K-NN successfully separates these clusters effectively, it indicates that the data clustering we performed is of good quality.

After the data was clustered by the K-Means algorithm (whose centroids were optimized by WOA), each data point was assigned a cluster label. We then used these cluster labels as class labels when training the K-NN algorithm. Thus, the original features of the data became the input for K-NN, and the previously generated cluster labels became the correct answers that K-NN learned.

To ensure that the K-NN testing results were accurate and not merely coincidental on a subset of the data, we used a cross-validation technique called Stratified K-Fold. This technique divides the data into several parts (folds), and the K-NN model is trained and tested alternately on these different parts. We chose Stratified K-Fold because this technique ensures that each part of the data has a balanced representation of each data cluster. This is crucial if the sizes of the data clusters we generated vary, so that the testing results remain fair and reliable.

The performance results of K-NN, especially its accuracy score (how often K-NN correctly predicts the data cluster), were used as an indicator of the data clustering quality. If the K-NN accuracy is high, it means that the data within one cluster tends to be similar and can be easily distinguished from the data in other clusters based on the existing features. This demonstrates that the data clustering we performed successfully formed clear and distinct groups.data.

3. Methods

In the next section, a detailed explanation will be provided regarding the literature review, research methodology, including the implementation of the WOA and K-NN algorithms, as well as the evaluation metrics used. Subsequently, the experimental results on various datasets and an analysis of those results will be presented.

Experimental Datasets

In this research, we used several datasets consisting of various data with multiple features[15]. These datasets were chosen due to their relevance to the research domain, sufficiently large dataset size, and the availability of class labels for evaluation[15]. This dataset has undergone an initial preprocessing process, such as the types of preprocessing performed, e.g., handling missing values, normalization, and outlier detection.

Table 1. Dataset Characteristic

Name	Dataset Characteristic		
	Number of Fitur	***Number of Class***	***Number of data***
Iris	4	3	150
Wine	13	3	178
Heart	13	2	303
Lung	15	2	309
Liver	10	2	583

3.2. Experimental Setup

This section will detail the Experimental Setup that we used to evaluate the WOA and K-NN integration approach in the clustering task. The WOA parameters (number of whales = 10, iterations= 100, and bounds) are determined based on the data. The determination of the number of K-Means clusters (number of clusters *k*) is based on the number of classes that each dataset possesses and various *k* values are not tested. Number of clusters k The K-NN parameters (number of neighbors= 1, distance metric= Euclidean distance) are also

specified along with the reasoning. The cross-validation configuration (number of folds=10) is explained. We use StrafiedKFold for better distribution in fold. The evaluation metrics for clustering (e.g., Davies-Bouldin) and K-NN validation (e.g., accuracy) are mentioned. The comparison method with a baseline algorithm (e.g., standard K-Means) is described. Finally, the computational environment and software (e.g., Python, scikit-learn) used are also informed

3.3. K-NN as a Validation Method

K-Nearest Neighbor (K-NN) is a simple yet effective classification algorithm [14]. K-NN works by classifying a data point based on the labels of its K nearest neighbors. In the context of clustering, K-NN can be used to validate the clustering results [15] in the following way:
1. Labeling Cluster: Each cluster is labeled based on the majority class of the data contained within it.
2. Test Data Classification: Test data is classified using the K-NN algorithm based on the predetermined cluster labels.
3. Accuracy Evaluation: Classification accuracy can be used as a metric to measure the quality of clustering.

3.4. Evaluation Metrics: Davies-Bouldin Index (DBI) and Accuracy

Evaluation metrics are very important for measuring the quality of clustering[16]. The Davies-Bouldin Index (DBI) is one of the commonly used metrics for evaluating clustering. DBI measures how well the resulting clusters are separated from each other and how compact each cluster is[17], [18]. A low DBI value indicates good clustering quality. A low DBI value indicates that the formed clusters are far apart from each other and have homogeneous members.

Besides DBI, accuracy can also be used as an evaluation metric, especially when the actual class labels of the data are known. Accuracy measures the proportion of data correctly classified by the K-NN algorithm after the cluster labeling process[15]. In the context of this research, both DBI and accuracy will be used to evaluate the performance of the proposed algorithm. Accuracy is the proportion of data that is classified correctly. In the context of clustering, accuracy is calculated by comparing the cluster labels given by the algorithm with the actual class labels of the data[14]

4. Results and Discussion

4.1. Comparison of Clustering Performance (DBI) with and Without WOA

This section will present the tangible outcomes of our clustering experiments. The primary emphasis lies in the quantitative exposition comparing the performance of the K-Means algorithm when initialized with centroids optimized by the Whale Optimization Algorithm (WOA) against standard K-Means employing random centroid initialization. For this comparison, we will leverage pertinent clustering evaluation metrics, and within the context of our discussed code, the Davies-Bouldin Index (DBI) will serve as a key indicator. It is crucial to underscore that a lower DBI value signifies superior clustering quality, characterized by more cohesive and well-separated clusters. We will juxtapose the DBI values obtained for both approaches (WOA-KMeans and standard K-Means) side-by-side, allowing for a clear discernment of the potential performance gains afforded by WOA integration.

Table 2. Davies Bouldin Index (DBI)

Dataset	DBI K-Means	
	Without WOA	***With WOA***
Iris	0,66604	0,66197
Wine	0,54956	0,53420
Heart	0,97701	0,96768
Lung	2,38016	0,65395
Liver	0,72101	0,62949

Beyond the quantitative results, it can be seen that the Davies-Bouldin Index (DBI) value is significantly lower in the model with WOA. A lower DBI value indicates that the resulting clusters are more compact and well-separated. To visualize the research data in the table, it can be represented in the **Fig. 1**:

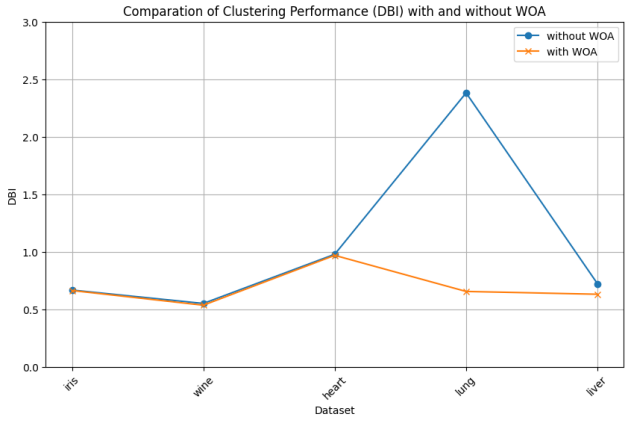

Fig. 1. Comparation of Clastering Performance (DBI) with and without WOA

4.2. Validation Results Using K-NN

This section will present the results of the K-NN classification performance, where the cluster labels generated by the K-Means algorithm (both with standard centroid initialization and those optimized by WOA) are utilized as the target variable. The primary objective here is to evaluate how well these cluster labels can be predicted by the K-NN model based on the original data features. The performance of K-NN will be assessed using relevant classification metrics.

Subsequently, we will conduct a comparison of the accuracy (and other classification metrics) obtained from the K-NN models trained on the cluster labels generated by both approaches: standard K-Means and WOA-optimized K-Means. This comparison will allow us to determine whether centroid optimization using WOA not only improves internal clustering metrics (such as DBI) but also yields more consistent and predictable cluster labels for an external classification model like K-NN.

Following this, we will perform an interpretation of the K-NN results within the context of clustering quality. The central argument here is that higher K-NN accuracy indicates that the resulting clusters are more distinct and meaningful. If the K-NN model can easily and accurately predict cluster labels from the data features, it implies that data points within the same cluster share similar feature characteristics and differ significantly from data points in other clusters. In other words, the decision boundaries learned by K-NN effectively separate the different clusters

Finally, we will engage in a discussion on how K-NN validation provides a different perspective on cluster quality compared to internal metrics. Internal metrics like DBI evaluate clustering quality based on the intrinsic characteristics of the clusters themselves (e.g., density and separation based on distances between points and centroids). Conversely, K-NN validation offers an external perspective by testing how "meaningful" the generated cluster labels are in the context of a classification task. If the cluster labels can be used to train a good classification model, it suggests that the clusters capture significant patterns in the data that can be generalized. The differences and complementarities between these two types of metrics will be discussed to provide a more holistic understanding of the achieved clustering quality.

The results of the K-NN classification performance is shown in the **Table 3** and **Table 4**. Average accuracy values derived from stratified K-fold cross-validation with k=10 are presented in **Table 3**. The standard deviations of the aforementioned accuracies are presented in **Table 4**.

Table 3. Average Accuracy

Dataset	K-NN Classfication performance with WOA (Average Accuracy)	
	Using K-Means Label	***Using True Label***
Iris	0,9800	0,9600
Wine	0,9888	0,7301
Heart	0,9900	0,6134
Lung	0,9871	0,8869
Liver	0,9983	0,6586

To visualize the research data in the **Table 3**, it can be represented in the **Fig. 2**:

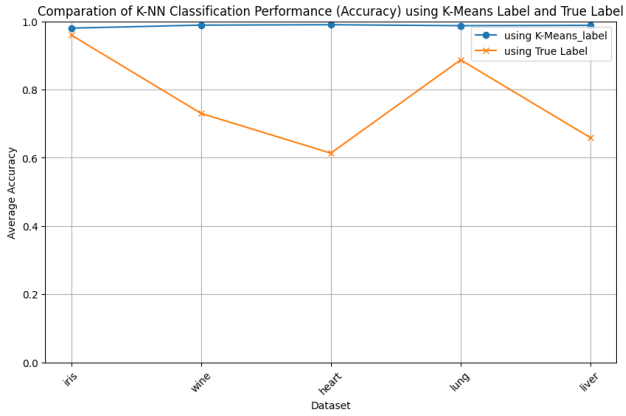

Fig. 2. Comparation of K-NN Classification Performance (Accuracy) using K-Means Label and True Label

Table 4. Standard Deviation of Accuracy

Dataset	K-NN Classfication performance with WOA (Standard Deviation Accuracy)	
	Using K-Means Label	***Using True Label***
Iris	0,03055	0,96000
Wine	0,03333	0,06079
Heart	0,01528	0,08227
Lung	0,01580	0,04777
Liver	0,00517	0,03959

4.3. Discussion

The DBI results indicate that the integration of WOA generally improved the clustering quality of K-Means (lower DBI) on the wine, heart, lung, and liver datasets, with a significant improvement observed on the lung dataset, where the DBI value drastically decreased from 2.38016 without WOA to 0.65395 with WOA. This substantial reduction suggests that WOA effectively assists K-Means in finding far better centroid configurations, resulting in significantly denser and more separated clusters within the feature space of the 'lung' dataset.

On the iris dataset, WOA did not yield a significant change. This may indicate that the 'iris' dataset possesses a relatively simple cluster structure that is easily discoverable even with random initialization, thus the benefit of centroid optimization by WOA is less pronounced in this case. However, it is important to note that WOA did not worsen the performance of K-Means on the 'iris' dataset.

The success of WOA is likely due to its ability to perform a better global search, avoiding local optima that often trap K-Means with random initialization. The exploration and exploitation phases of WOA enable a more effective search of the centroid space. The DBI as an internal metric supports the improvement in cluster quality; however, external validation (such as K-NN) is necessary for the perspective of cluster meaningfulness. Future research should compare these findings with similar studies.

The advantage of WOA-KMeans is its potential to enhance clustering quality, especially on complex data. Its limitation is the additional computational cost. In big data analysis, this approach has the potential to yield better and more stable clustering. Generalization to large and high-dimensional data requires further research, including the sensitivity of WOA and K-Means parameters.

Furthermore, external validation using K-NN classification with K-Means cluster labels as the target variable yielded very high accuracy (98-99%) across all datasets. This implies that the formed clusters, regardless of the centroid initialization method (with or without WOA in these K-NN accuracy results), exhibit good internal homogeneity and inter-cluster separation within the feature space. The ability of K-NN to accurately predict

cluster labels confirms that the clustering algorithm successfully identified distinct groups of data based on their feature characteristics.

Conversely, the performance of K-NN classification using the true class labels of the datasets showed greater variation and generally lower accuracy, particularly on the wine, heart, and liver datasets. This disparity indicates that the clusters formed by K-Means, being an unsupervised algorithm, do not always align with the existing ground truth class divisions. However, the high K-NN accuracy with cluster labels remains relevant as it validates the existence of distinct and separated structures in the data based on feature similarity.

5 Conclusion

Overall, the DBI and K-NN accuracy results provide complementary perspectives on clustering quality. The decrease in DBI with WOA integration indicates an improvement in internal cluster cohesion and separation. Meanwhile, the high K-NN accuracy in predicting cluster labels confirms the external validity of the formed clusters based on separability in the feature space. Even though the generated clusters do not always reflect the true class labels, the success of K-NN in classifying based on cluster labels indicates that the clustering algorithm, especially when initialized with WOA, is capable of capturing meaningful patterns and structures within the data. Future research can compare these findings with other studies using similar approaches and further explore the influence of WOA and K-Means parameters.

Acknowledgment

The authors would like to express their sincere gratitude the research assistants for their diligent data collection and their motivation and the Muhammadiyah University of Brebes for providing computational resources.

The authors also acknowledge the availability of open-source software libraries, particularly scikit-learn in Python, which proved invaluable in the implementation and evaluation of the algorithms used in this study.

Finally, the authors appreciate the time and effort of the anonymous reviewers whose constructive feedback helped to improve the quality and clarity of this manuscript.

References

[1] G. J. Oyewole and G. A. Thopil, “Data clustering: application and trends,” *Artif Intell Rev*, vol. 56, no. 7, pp. 6439–6475, Jul. 2023, doi: 10.1007/s10462-022-10325-y.
[2] T. Dinh *et al.*, “Data clustering: an essential technique in data science,” 2024, *arXiv*. doi: 10.48550/ARXIV.2412.18760.
[3] V. V. Baligodugula and F. Amsaad, “Unsupervised Learning: Comparative Analysis of Clustering Techniques on High-Dimensional Data,” 2025, *arXiv*. doi: 10.48550/ARXIV.2503.23215.
[4] J. Nasiri and F. M. Khiyabani, “A whale optimization algorithm (WOA) approach for clustering,” *Cogent Mathematics & Statistics*, vol. 5, no. 1, p. 1483565, Jan. 2018, doi: 10.1080/25742558.2018.1483565.
[5] M. Ahmed, R. Seraj, and S. M. S. Islam, “The k-means Algorithm: A Comprehensive Survey and Performance Evaluation,” *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020, doi: 10.3390/electronics9081295.
[6] H. Singh *et al.*, “An enhanced whale optimization algorithm for clustering,” *Multimed Tools Appl*, vol. 82, no. 3, pp. 4599–4618, Jan. 2023, doi: 10.1007/s11042-022-13453-3.
[7] S. Mirjalili and A. Lewis, “The Whale Optimization Algorithm,” *Advances in Engineering Software*, vol. 95, pp. 51–67, May 2016, doi: 10.1016/j.advengsoft.2016.01.008.
[8] D. Liauw, M. Q. Khairuzzaman, and G. Syarifudin, “Whale Optimization Algorithm for Data Clustering,” in *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, Jakarta, Indonesia: IEEE, Nov. 2019, pp. 1–6. doi: 10.1109/CITSM47753.2019.8965415.
[9] P. K. Syriopoulos, N. G. Kalampalikis, S. B. Kotsiantis, and M. N. Vrahatis, “kNN Classification: a review,” *Ann Math Artif Intell*, vol. 93, no. 1, pp. 43–75, Feb. 2025, doi: 10.1007/s10472-023-09882-x.
[10] M. Suyal and P. Goyal, “A Review on Analysis of K-Nearest Neighbor Classification Machine Learning Algorithms based on Supervised Learning,” *IJETT*, vol. 70, no. 7, pp. 43–48, Jul. 2022, doi: 10.14445/22315381/IJETT-V70I7P205.
[11] M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs, “**NbClust** : An *R* Package for Determining the Relevant Number of Clusters in a Data Set,” *J. Stat. Soft.*, vol. 61, no. 6, 2014, doi: 10.18637/jss.v061.i06.
[12] Y. Zhou and Z. Hao, “Multi-Strategy Improved Whale Optimization Algorithm and Its Engineering Applications,” *Biomimetics*, vol. 10, no. 1, p. 47, Jan. 2025, doi: 10.3390/biomimetics10010047.

[13] N. W. Hidayat, . Purwanto, and F. Budiman, "Whale Optimization Algorithm Bat Chaotic Map Multi Frekuensi for Finding Optimum Value," *JAIS*, vol. 5, no. 2, pp. 80–90, Feb. 2021, doi: 10.33633/jais.v5i2.4432.
[14] L. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, p. 1883, 2009, doi: 10.4249/scholarpedia.1883.
[15] S. Hulu and P. Sihombing, "Analysis of Performance Cross Validation Method and K-Nearest Neighbor in Classification Data," *International Journal of Research and Review*, no. 4, 2020.
[16] D. L. Davies and D. W. Bouldin, "A Cluster Separation Measure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.
[17] Y. Arie Wijaya, D. Achmad Kurniady, E. Setyanto, W. Sanur Tarihoran, D. Rusmana, and R. Rahim, "Davies Bouldin Index Algorithm for Optimizing Clustering Case Studies Mapping School Facilities," *TEM Journal*, pp. 1099–1103, Aug. 2021, doi: 10.18421/TEM103-13.
[18] A. Idrus, N. Tarihoran, U. Supriatna, A. Tohir, S. Suwarni, and R. Rahim, "Distance Analysis Measuring for Clustering using K-Means and Davies Bouldin Index Algorithm," *TEM Journal*, pp. 1871–1876, Nov. 2022, doi: 10.18421/TEM114-55.

.