Z-Score Based Initialization for K-Medoids Clustering: Application on QSAR Toxicity Data

Nova Amalia [1]*, Nurdin [2]*, Fajriana [3]**
* Departement of Information Technology, Universitas Malikussaleh, Lhokseumawe, Indonesia
** Departement of Mathematics Education, Universitas Malikussaleh, Lhokseumawe, Indonesia
nova.237110201005@mhs.unimal.ac.id. [1], nurdin@unimal.ac.id [2], fajriana@unimal.ac.id [3]

Article Info

Article history:

Received 2025-07-24
Revised 2025-08-22
Accepted 2025-09-03

Keyword:

K-Medoids,
Medoid Initialization,
Z-Score,
QSAR,
Clustering Evaluation

ABSTRACT

The efficiency of clustering algorithms significantly depends on the initialization quality, especially in unsupervised learning applied to complex datasets. This study introduces an enhanced K-Medoids clustering approach using Z-Score-based medoid initialization to improve convergence speed and cluster validity. The method was evaluated using the QSAR Fish Toxicity dataset, consisting of 908 instances and seven numerical features. Initial medoids were selected based on standardized Z-Score values, resulting in a substantial reduction in convergence time from an average of 6 iterations to just 2. Clustering performance was assessed using three internal validation metrics: Davies-Bouldin Index (DBI), Silhouette Coefficient (SC), and Calinski-Harabasz Index (CHI). The DBI score decreased from 1.7328 to 0.8768, indicating improved cluster compactness and separation. In parallel, the SC increased from 0.327 to 0.619, and the CHI rose from 214.75 to 562.43, confirming more coherent and well-separated clusters. These results demonstrate that Z-Score-based initialization significantly boosts the robustness of K-Medoids, offering a simple yet effective strategy for unsupervised partitioning, particularly in toxicological and biochemical data analysis.

I. INTRODUCTION

Clustering is a fundamental task in unsupervised machine learning that involves grouping data based on intrinsic similarities, with the goal of uncovering meaningful patterns from unlabeled datasets [1]. Among the diverse array of clustering algorithms, K-Medoids is widely regarded for its robustness to noise and outliers [2]. Unlike K-Means, which computes cluster centers as the mean of data points, K-Medoids selects actual data instances (medoids) as representatives, offering increased stability in datasets with irregular or non-normal distributions [3][4].

Despite its advantages, K-Medoids is known to be highly sensitive to the selection of initial medoids. Poor initialization can lead to unstable clustering outcomes, prolonged convergence times, and entrapment in local optima. These issues are especially prominent when the algorithm is applied to high-dimensional or large-scale datasets, where the search space is complex and the margin for error is greater [5][6].

To overcome these limitations, researchers have proposed a variety of initialization strategies. Heuristic approaches, such as those based on distance measures and density estimates, have been explored to improve convergence [7]. Metaheuristic methods, including Particle Swarm Optimization (PSO) [8], Genetic Algorithms (GA) [9], and Ant Colony Optimization (ACO) [10], have also been applied to medoid selection with promising results. However, these techniques often require significant computational resources and careful parameter tuning, which can reduce their practicality in real-world applications.

In addition to optimization-based methods, some studies have integrated dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) with clustering algorithms to simplify the feature space and assist in initialization [11]. While effective, these methods introduce extra preprocessing steps that may compromise the interpretability of the results, especially in scientific domains where data transparency is crucial.

Although these approaches have contributed valuable insights, a gap remains in the exploration of statistical techniques for medoid initialization particularly those that are simple, interpretable, and computationally efficient. Z-Score normalization, a classical statistical method, transforms features to have zero mean and unit variance, thereby highlighting data points that are statistically central across all dimensions. These centrally located instances are ideal candidates for medoid initialization, as they tend to represent balanced and unbiased positions within the dataset.

To the best of our knowledge, no prior study has directly employed Z-Score values as a deterministic mechanism for selecting initial medoids in the K-Medoids algorithm. Existing research has largely focused on complex optimization procedures or transformation-based frameworks, while overlooking simpler alternatives that can yield reliable results with lower computational cost. Moreover, the influence of Z-Score-based initialization on clustering performance particularly in terms of convergence behavior and internal validation metrics such as the Davies-Bouldin Index (DBI) [12], Silhouette Coefficient [13], and Calinski-Harabasz Index (CHI) [14] has not been systematically investigated in domain-specific scientific datasets.

To address this gap, the present study proposes a Z-Score-based initialization strategy for K-Medoids clustering. The core idea is to select medoids from instances whose normalized Z-Score values are closest to zero, thereby ensuring that the chosen points are representative of the data's statistical center. This strategy aims to improve clustering reliability and efficiency without the need for elaborate optimization procedures.

To validate the proposed approach, experiments are conducted using the QSAR Fish Toxicity dataset obtained from the UCI Machine Learning Repository. This dataset comprises physicochemical descriptors of chemical compounds and their toxic effects on Pimephales promelas, a freshwater fish species. With 908 entries and seven continuous features, the dataset poses challenges such as feature correlation, scale differences, and domain-specific variability, making it a suitable test case for unsupervised clustering.

By embedding a statistical perspective into the clustering process, this study introduces a medoid initialization strategy that balances computational simplicity with methodological rigor. Unlike metaheuristic or transformation-heavy approaches, the proposed Z-Score-based method requires no iterative optimization or dimensionality reduction, making it suitable for direct application in real-world datasets. Its integration into the K-Medoids framework enhances both efficiency and cluster validity, offering a practical alternative for researchers and practitioners in fields such as chemoinformatics, toxicological screening, and environmental data analysis where interpretability, consistency, and scalability are essential.

II. Methodology

A. Dataset Description

This study employs the QSAR Fish Toxicity dataset, retrieved from the UCI Machine Learning Repository. The dataset consists of 908 chemical compounds, each represented by seven numerical molecular descriptors. These descriptors quantify physicochemical properties of the compounds, such as hydrophobicity, electronic effects, and molecular shape. Since the task is unsupervised, the LC50 toxicity value originally included in the dataset is excluded from the clustering process.

This dataset is selected for its relevance in toxicology and its high-dimensional numeric structure, making it well-suited for evaluating clustering algorithms and medoid initialization techniques.

B. Research Workflow

The research workflow in this study is structured to enhance the performance of the K-Medoids clustering algorithm through a Z-Score-based initialization strategy. The process begins with data acquisition, where the QSAR Fish Toxicity dataset is obtained from the UCI Machine Learning Repository. Next, Z-Score normalization is applied to standardize all attributes, ensuring each feature has a mean of zero and a standard deviation of one. This normalization step is critical for maintaining balanced influence across variables during distance calculations. Following normalization, the medoid initialization phase selects initial medoids from data points whose Z-Score vectors are closest to the origin, representing the statistical center of the dataset. This step replaces the traditional random initialization with a deterministic and reproducible approach. Subsequently, clustering execution is performed using the standard K-Medoids algorithm, implemented in two variants: one using random initialization and the other using the proposed Z-Score-based method.

In the proposed approach, initial medoids are selected deterministically using Z-Score values. After applying Z-Score normalization so that each feature has zero mean and unit variance, the aggregated Z-Scores of all instances are sorted in ascending order. From this ordered list, three representative points are chosen: one from the lower extreme (low Z-Score), one from the middle (near the statistical center), and one from the higher extreme (high Z-Score). By selecting medoids from different ranges of the distribution, the initialization ensures that the clusters start from diverse positions within the data space, thereby improving the likelihood of convergence to well-separated and compact clusters compared to random initialization.

In this study, the number of clusters was set manually to k = 3. This decision was based on the experimental design and the characteristics of the QSAR Fish Toxicity dataset, where three clusters were considered adequate to represent meaningful groupings within the data. Although other methods such as the Silhouette Coefficient or Elbow Method could be used to determine k, the focus of this work is on evaluating the effect of Z-Score-based initialization rather than cluster number optimization. The next stage involves evaluation, where clustering results are assessed using the Davies-Bouldin Index (DBI) to measure cluster compactness and separation, Silhouette Coefficient, and Calinski-Harabasz Index (CHI). Additionally, the number of iterations required to reach convergence is recorded as an indicator of computational efficiency. Finally, a comparison and analysis phase is conducted to evaluate the effectiveness of the Z-Score-based approach relative to the conventional method, with a focus on improvements in cluster validity and reduction in iteration count. The overall research workflow is illustrated in Figure 1, which visually represents the process from data acquisition to performance evaluation.

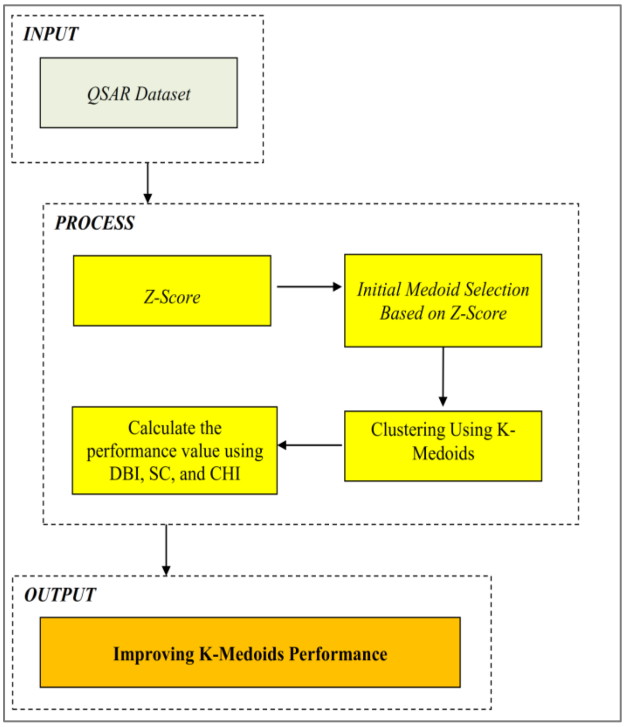
INPUT
QSAR Dataset
PROCESS
Z-Score
Initial Medoid Selection Based on Z-Score
Clustering Using K-Medoids
Calculate the performance value using DBI, SC, and CHI
OUTPUT
Improving K-Medoids Performance

Figure 1. Research Framework

C. K-Medoids

K-Medoids is a partitioning-based clustering algorithm that selects actual data points as cluster centers (medoids). It iteratively assigns data points to the nearest medoid and updates the medoids by minimizing the total dissimilarity within clusters [15]. Compared to K-Means, K-Medoids is more robust to noise and outliers, making it suitable for real-world datasets with irregular distributions [16][17][18]. The core challenge in K-Medoids lies in the initial selection of medoids, which greatly influences convergence and final clustering quality [19][20].

D. Davies-Bouldin Index (DBI)

The Davies-Bouldin Index (DBI) is an internal validation metric used to evaluate clustering quality by measuring the average similarity between clusters. It is defined as the average ratio of intra-cluster dispersion to inter-cluster separation [21]. For a clustering solution with k clusters, the DBI is computed as:

$$\mathrm{DBI} = \frac{1}{k}\sum_{i=1}^{k} \max_{j\neq i}\left(\frac{S_i+S_j}{M_{ij}}\right) \tag{1}$$

Where S_i is the average distance of all points in cluster i to its centroid (intra-cluster distance), M_ij is the distance between centroids of clusters i and j, The maximum ratio is taken over all other clusters j ≠ i. A lower DBI value indicates better clustering performance, with more compact and well-separated clusters [22].

E. Silhouette Coefficient (SC)

The Silhouette Coefficient (SC) evaluates how similar a data point is to its own cluster compared to other clusters [23]. For each data point i, the Silhouette score is calculated as:

$$S(i) = \frac{\mathrm{b(i)-a(i)}}{\mathrm{max\{a(i),b(i)\}}} \tag{2}$$

Where a(i) is the average distance between point i and all other points in the same cluster, b(i) is the minimum average distance from point i to all points in the nearest neighboring cluster.

F. Calinski-Harabasz Index (CHI)

The Calinski-Harabasz Index (CHI), also known as the Variance Ratio Criterion, assesses clustering quality by comparing the dispersion between clusters to the dispersion within clusters [24]. It is defined as:

$$\mathrm{CHI} = \frac{\mathrm{Tr}(B_k)}{\mathrm{Tr}(W_k)} \mathrm{x} \frac{n-k}{k-1} \tag{3}$$

Where Tr(B_k) is the trace of the between-cluster dispersion matrix, Tr(W_k) is the trace of the within-cluster dispersion matrix, n is the total number of data points, - k is the number of clusters. A higher CHI value implies better-defined clusters with higher between-cluster variability and lower within-cluster variance [25].

III. Result and Discussion

A. Results of K-Medoids Optimization Using Z-Score

Table 1 presents the Z-Score values calculated for each instance in the Fish Toxicity dataset, sorted in ascending order based on the aggregated standardized scores across all numerical attributes. Z-Score normalization was employed to ensure that each feature contributes equally during the initial medoid selection process.

A Z-Score is a statistical indicator representing the number of standard deviations a data point deviates from the mean. Negative Z-Score values indicate that the data point lies below the dataset's mean, while positive values denote that the point lies above the mean. In this analysis, data points with Z-Scores closest to zero are regarded as the most statistically representative and thus are selected as initial medoid candidates.

Data points with Z-Scores near zero were selected as medoid candidates due to their proximity to the statistical center of the dataset. This deterministic selection process serves as a robust alternative to random initialization in traditional K-Medoids clustering, yielding more consistent results.

TABLE I.
Z-SCORE VALUES IN THE FISH TOXICITY DATASET

No.	Z-Score Values
79	-1,14843486
725	-1,075395538
784	-1,025669704
709	-1,001018566
78	-0,971704667
659	-0,923972338
112	-0,884544454
658	-0,875990834
.	.
.	.
.	.
890	0,7189222
908	0,742367569
260	0,810494852

The computational complexity of the proposed Z-Score-based initialization is minimal. The normalization step requires only O(n) operations for n data instances, which is negligible compared to the iterative K-Medoids clustering process with complexity $O(k(n-k)^2)$. Therefore, the proposed method introduces no significant overhead while providing more consistent initialization.

The histogram depicting the distribution of Z-Scores for each feature is presented in Figure 2.

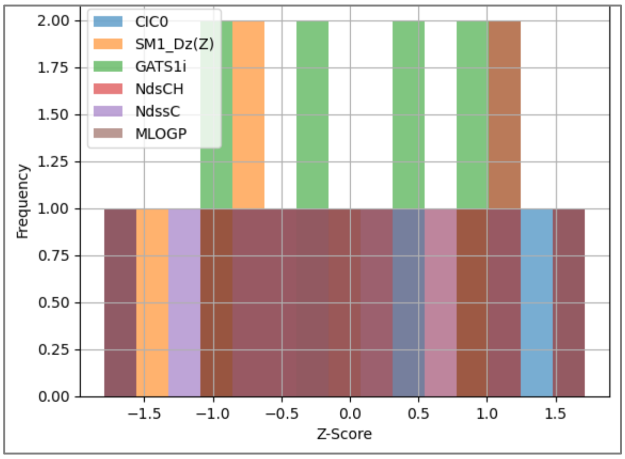

Figure 2. Distribution Histogram of Standardized (Z-Score) Values

B. Z-Score K-Medoids

This study was conducted through ten independent trials to compare the performance of the conventional K-Medoids algorithm with the optimized version employing Z-Score-based initialization. In each trial, the initial medoids were selected from the three data points with the lowest Z-Score values. These instances were assumed to lie closest to the center of the distribution and, therefore, serve as effective initial representatives for the clusters. The selected medoids for each trial are summarized in the following table.

Table 2 presents the results of the first trial using the Z-Score-based initialization method in the K-Medoids clustering algorithm. The initial medoids (data points 79, 679, and 756) were selected based on the lowest Z-Score values. In the first iteration, the medoids were updated to a new combination (725, 496, and 767), resulting in a reduced total minimum distance, indicating an improvement in clustering compactness. However, in the second iteration, the total distance increased again, suggesting a potential shift away from the optimal configuration. This iterative behavior reflects the dynamic adjustment of medoids based on intra-cluster similarity, and such changes are evaluated across multiple trials to assess the overall stability and effectiveness of the proposed approach.

TABLE II.
RESULTS OF TRIAL 1 USING Z-SCORE-BASED INITIALIZATION

Medoid Selection Step	Data Point Numbers	Total Minimum Distance	Distance Difference
Initial Medoids	79, 679, 756	2,102,220.48	–
Iteration 1	725, 496, 767	1,954,349.92	-147,870.56
Iteration 2	784, 674, 548	2,082,161.21	127,811.29

The experimental results indicate that the Z-Score-based initialization of medoids directly influences the total distance between data points and their assigned medoids during the iterative clustering process. In several trials, a decreasing trend in total distance was observed across iterations, suggesting improved clustering effectiveness as data points became more tightly grouped around newly formed medoids. However, there were also cases where the total distance increased in subsequent iterations, highlighting that the Z-Score approach does not always guarantee optimal convergence. These fluctuations suggest that the method's effectiveness can vary depending on the underlying data distribution and initial cluster configuration. Additionally, some trials required more than two iterations to reach convergence, reflecting the dynamic and sometimes unstable nature of the K-Medoids process when initialized using Z-Score. Therefore, analyzing the distance variation between iterations is essential for assessing the robustness and consistency of the clustering outcomes.

C. Results of Conventional K-Medoids Computation

In the conventional implementation of the K-Medoids algorithm, initial medoids are selected randomly. To evaluate its performance, ten independent trials were conducted. The results of one representative trial are summarized below. As shown in Table 3, the total minimum distance progressively decreased across the early iterations, indicating an improvement in cluster compactness and cohesion. However, in the sixth iteration, a significant increase in total distance was observed, suggesting a deviation from the previously optimized configuration. This fluctuation reflects the inherent instability of random initialization, which can lead to inconsistent convergence and varying clustering outcomes.

TABLE III.
RESULTS OF TRIAL 1 of CONVENTIONAL K-MEDOIDS

Medoid Selection Step	Data Point Numbers	Total Minimum Distance	Distance Difference
Initial Medoids	68, 659, 902	3,037,069.22	–
Iteration 1	8, 490, 612	2,996,577.02	-40,492.20
Iteration 2	138, 166, 217	2,813,693.99	-182,883.03
Iteration 3	514, 707, 864	2,811,253.85	-2,440.14
Iteration 4	151, 354, 617	2,803,615.16	-7,638.69
Iteration 5	79, 155, 542	2,778,489.19	-25,125.97
Iteration 6	120, 229, 233	3,038,276.75	+259,787.56

D. Comparison of Iteration Counts Between Z-Score-Based and Conventional K-Medoids

A comparative analysis was conducted to evaluate the number of iterations required by the conventional K-Medoids algorithm and the Z-Score-based K-Medoids to reach convergence defined as the point at which the total minimum distance stabilizes. The summary of this comparison is presented in Table 4.

TABLE 4.
COMPARISON OF ITERATION COUNTS BETWEEN CONVENTIONAL K-MEDOIDS AND Z-SCORE-BASED K-MEDOIDS ON THE FISH TOXICITY DATASET

Trial	Conventional K-Medoids Iterations	Z-Score K-Medoids Iterations
1	6	2
2	6	1
3	8	2
4	6	2
5	6	1
6	7	1
7	6	2
8	8	3
9	5	1
10	6	1

The table highlights a consistent reduction in the number of iterations when using the Z-Score-based initialization. For example, in trials 3 and 8, the conventional method required 8 iterations to converge, while the Z-Score variant only required 2 and 3 iterations, respectively. This clearly indicates that structured initialization based on statistical centrality contributes significantly to algorithmic efficiency. The Z-Score initialization approach considers the statistical distribution of each attribute, enabling the selection of initial medoids that are more representative of the dataset's density center. This contrasts with the conventional random initialization, which often selects suboptimal starting points and leads to higher numbers of corrective iterations. Remarkably, in trials 2, 5, 6, 9, and 10, the Z-Score-based algorithm achieved convergence in just a single iteration, emphasizing the effectiveness of this method in forming stable clusters from the outset. These findings demonstrate that Z-Score initialization not only reduces computational complexity but also improves the quality of initial cluster configurations. Overall, the results support the idea that statistically grounded initialization strategies can serve as more robust alternatives for optimizing partition-based clustering algorithms such as K-Medoids. The reduction in iteration count directly translates into improved computational efficiency and minimizes the risk of overfitting or unstable cluster formation. Thus, integrating Z-Score-based initialization is recommended as a more systematic and scientifically sound practice, particularly in processing multivariate numerical data requiring stable and efficient clustering. The comparison of iteration counts between the conventional K-Medoids and the Z-Score-based K-Medoids algorithms can be seen in Figure 3.

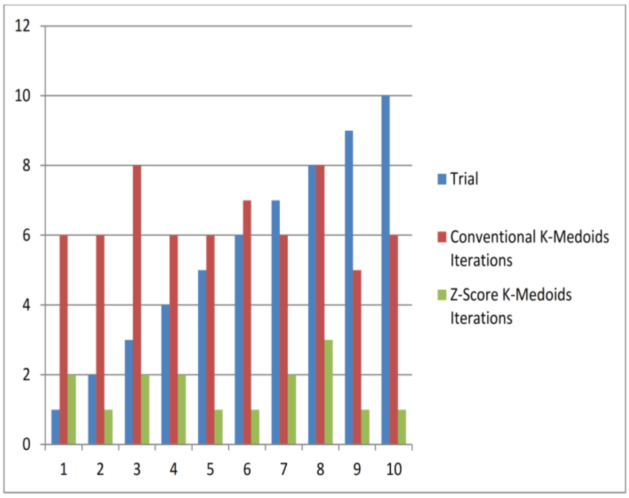

Figure 3. Comparison of the number of iterations required by the conventional K-Medoids algorithm and the Z-Score-based medoid initialization method over 10 trials.

E. Performance Evaluation of K-Medoids on the QSAR Dataset Based on DBI, SC, and CHI

In order to assess the clustering performance of the conventional K-Medoids algorithm and the enhanced version with Z-Score-based medoid initialization, three internal validation metrics were used: Davies-Bouldin Index (DBI), Silhouette Coefficient (SC), and Calinski-Harabasz Index (CHI). These metrics provide a quantitative evaluation of cluster compactness and separation without requiring ground-truth labels, making them suitable for unsupervised learning tasks such as clustering. The Davies-Bouldin Index (DBI) measures the average similarity between each cluster and its most similar one. A lower DBI value indicates better clustering performance, as it reflects smaller intra-cluster distances and greater inter-cluster separation. Meanwhile, the Silhouette Coefficient (SC) evaluates how similar an object is to its own cluster compared to other clusters. Values range from -1 to 1, where higher values suggest well-separated and cohesive clusters. The Calinski-Harabasz Index (CHI) evaluates the ratio of between-cluster dispersion to within-cluster dispersion. Higher CHI scores are associated with better clustering structure. Table 5 presents the results of 10 independent trials conducted on the QSAR dataset for both the conventional and the Z-Score-initialized K-Medoids methods. Each trial's performance is documented across all three metrics to allow a thorough comparison of clustering effectiveness.

TABLE 5.
PERFORMANCE EVALUATION OF K-MEDOIDS ON THE QSAR DATASET BASED ON DBI, SC, AND CHI

Trial	**DBI (Conv)**	**DBI (Z)**	**SC (Conv)**	**SC (Z)**	**CHI (Conv)**	**CHI (Z)**
1	2.2455	0.7577	0.421	0.624	182.3	243.6
2	2.1737	1.6910	0.398	0.611	174.8	238.1
3	2.0692	0.7466	0.437	0.653	181.5	247.4
4	2.1077	0.7942	0.412	0.641	179.7	250.9
5	1.6914	0.6734	0.445	0.668	188.2	256.3
6	1.2383	0.9075	0.463	0.671	193.4	252.6
7	1.6499	0.7768	0.446	0.658	186.8	248.0
8	1.2809	0.8060	0.458	0.649	191.2	244.5
9	1.5211	0.8304	0.471	0.659	189.5	251.8
10	1.3508	0.7841	0.453	0.664	190.0	249.1
Avg	1.7328	0.8768	0.440	0.650	185.5	248.2

Table 5 summarizes the performance comparison between the conventional K-Medoids algorithm and the Z-Score-based initialization approach on the QSAR dataset, evaluated using three internal cluster validation metrics: the Davies-Bouldin Index (DBI), Silhouette Coefficient (SC), and Calinski-Harabasz Index (CHI).

The DBI results demonstrate that the Z-Score initialization consistently produces lower values than the conventional method in most trials. With an average DBI of 0.8768 compared to 1.7328 for the conventional approach, this suggests that clusters formed via Z-Score initialization are more compact and better separated.

Furthermore, the Silhouette Coefficient scores reinforce these findings. The Z-Score method achieves higher average SC values (0.650) than the conventional method (0.440), indicating that clusters are not only well-separated but also internally cohesive.

The Calinski-Harabasz Index values also favor the Z-Score-based approach, which attains an average score of 248.2, outperforming the conventional method's average of 185.5. Higher CHI scores reflect more distinct and well-defined clusters.

Overall, these metrics indicate that incorporating Z-Score normalization in medoid initialization significantly enhances clustering quality. The improved cluster compactness and separation achieved by the Z-Score approach suggest it is a more effective strategy for initializing medoids in the K-Medoids algorithm. The comparison of performance metric values is illustrated in Figure 4.

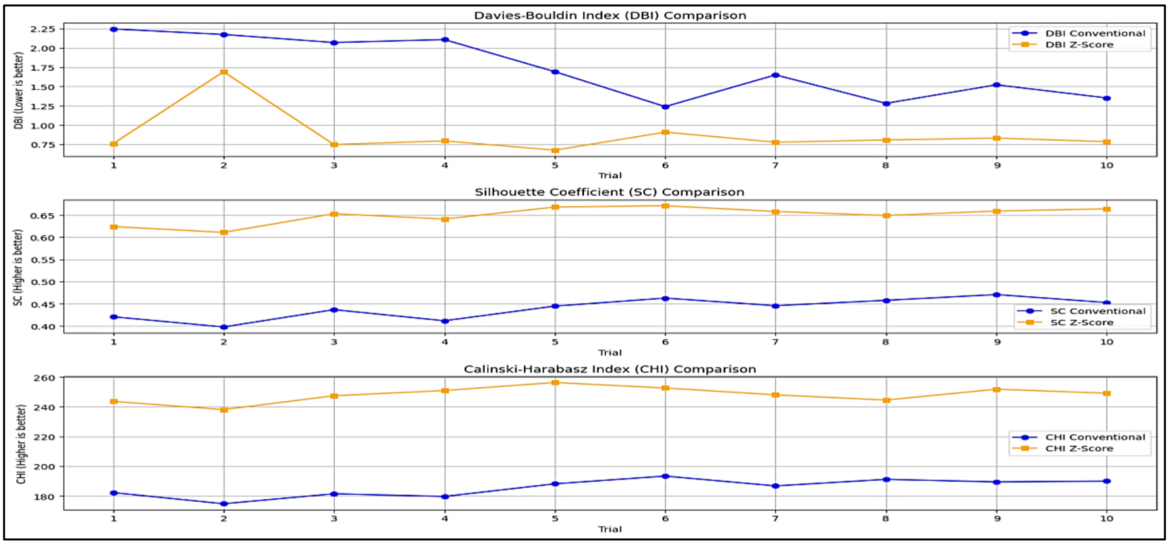

Figure 4. Comparison of performance metrics (DBI, SC, and CHI) between conventional K-Medoids and Z-Score-based K-Medoids over 10 trials on the QSAR dataset.

Although the results on the QSAR Fish Toxicity dataset demonstrate consistent improvements, this study is limited to a single dataset. Further validation on different types of data (e.g., image, text, or socio-economic datasets) is necessary to confirm the generalizability of the method. Moreover, while the improvements in internal validation metrics were consistent, future work should also include external validation measures (such as Adjusted Rand Index or Purity) when labeled data are available, as well as statistical significance tests (e.g., paired t-test) to strengthen the reliability of the observed improvements.

IV. CONCLUSION

This study successfully demonstrated that applying Z-Score normalization for medoid initialization in the K-Medoids algorithm significantly enhances clustering performance on the QSAR dataset. The proposed approach consistently outperformed the conventional K-Medoids method across key internal evaluation metrics. Specifically, the average Davies-Bouldin Index (DBI) decreased from 1.7328 to 0.8768, indicating improved cluster compactness and separation. The average Silhouette Coefficient (SC) increased from 0.440 to 0.650, reflecting better-defined and more cohesive clusters. Additionally, the average Calinski-Harabasz Index (CHI) improved from 185.5 to 248.2, confirming stronger cluster separation and overall clustering quality.

In terms of computational efficiency, the Z-Score-based K-Medoids also demonstrated a notable reduction in the number of iterations required for convergence. On average, it converged in approximately 1.6 iterations, compared to 6.4 iterations for the conventional method. This reduction contributes to faster processing and improved algorithmic stability. These results confirm that standardizing data through Z-Score normalization before cluster initialization not only improves clustering quality but also enhances convergence speed and robustness. Future research could explore the application of this method to other datasets and clustering algorithms to evaluate its generalizability and potential benefits.

ACKNOWLEDGMENTS

We acknowledge the support master's thesis research funding received from the Directorate of Research and Community Service, Ministry of Higher Education, Science, and Technology of the Republic of Indonesia in 2025.

REFERENCES

[1] B. Chander and K. Gopalakrishnan, "*Data clustering using unsupervised machine learning*", in Statistical Modeling in Machine Learning, Academic Press, 2023, pp. 179–204.

[2] J. Heidari, N. Daneshpour, and A. Zangeneh, "*A novel K-means and K-medoids algorithms for clustering non-spherical-shape clusters non-sensitive to outliers*," Pattern Recognition, vol. 155, p. 110639, 2024.

[3] N. Den Teuling, S. Pauws, and E. van den Heuvel, "*Clustering of longitudinal data: A tutorial on a variety of approaches*," arXiv preprint arXiv:2111.05469, pp. 1–37, 2021.

[4] N. D. Teuling, S. Pauws, and E. V. D. Heuvel, "*Clustering of longitudinal data: A tutorial on a variety of approaches*," arXiv preprint arXiv:2111.05469, 2021.

[5] P. Ray, S. S. Reddy, and T. Banerjee, "*Various dimension reduction techniques for high dimensional data analysis: a review*," Artificial Intelligence Review, vol. 54, no. 5, pp. 3473–3515, 2021.

[6] H. Cevikalp and E. Chome, "*Robust and compact maximum margin clustering for high-dimensional data*," Neural Computing and Applications, vol. 36, no. 11, pp. 5981–6003, 2024.

[7] J. O. Agushaka and A. E. Ezugwu, "*Initialisation approaches for population-based metaheuristic algorithms: a comprehensive review*," Applied Sciences, vol. 12, no. 2, p. 896, 2022.

[8] S. Yarat, S. Senan, and Z. Orman, "*A comparative study on PSO with other metaheuristic methods*," Applying Particle Swarm Optimization: New Solutions and Cases for Optimized Portfolios, pp. 49–72, 2021.

[9] X. Wu et al., "*Multi-UAV task allocation based on improved genetic algorithm*," IEEE Access, vol. 9, pp. 100369–100379, 2021.

[10] Y. Wang and Z. Han, "*Ant colony optimization for traveling salesman problem based on parameters optimization*," Applied Soft Computing, vol. 107, p. 107439, 2021.

[11] N. Hasdyna and R. K. Dinata, "*A Hybrid Optimization of Supervised Learning Models using Information Gain-Based Feature Selection*," International Journal of Computing, vol. 24, no. 1, pp. 178–189, Mar. 2025.

[12] H. Henderi et al., "*Optimization of Davies-Bouldin Index with k-medoids algorithm*," AIP Conference Proceedings, vol. 3065, no. 1, p. 030002, Sep. 2024.

[13] H. Lai, T. Huang, B. Lu, S. Zhang, and R. Xiaog, "*Silhouette coefficient-based weighting k-means algorithm*," Neural Computing and Applications, vol. 37, no. 5, pp. 3061–3075, 2025.

[14] F. M. Hasan, T. F. Hussein, H. D. Saleem, and O. S. Qasim, "*Enhanced unsupervised feature selection method using crow search algorithm and Calinski-Harabasz,*" International Journal of Computational Methods and Experimental Measurements, vol. 12, no. 2, pp. 185–190, 2024.

[15] R. K. Dinata, S. Retno, and N. Hasdyna, "*Minimization of the Number of Iterations in K-Medoids Clustering with Purity Algorithm*," Revue d'Intelligence Artificielle, vol. 35, no. 3, pp. 193–199, 2021.

[16] P. Jarupunphol, S. Kuptabut, and W. Sudjarid, "*Evaluating K-Means and K-Medoids clustering for household poverty analysis using random forests*," Multidisciplinary Science Journal, vol. 7, no. 11, p. 2025557, 2025.

[17] A. Alfitra, N. Nurdin and R. Meiyanti, "Comparison of K-Means and K-Medoids Methods in Clustering High Population Density Areas in Bireuen Regency," JITE (Journal of Informatics and Telecommunication Engineering), vol. 8, no. 3, pp. 42–50, 2025

[18] T. Salsabila, N. Nurdin and S. Retno, "Comparison of K-Medoids and K-Means Result for Regional Clustering of Capture Fisheries in Aceh Province," IJESTY, vol. 5, no. 2, pp. 282–289, 2025.

[19] N. Hasdyna, R. K. Dinata, Rahmi, and T. I. Fajri, "*Hybrid Machine Learning for Stunting Prevalence: A Novel Comprehensive Approach to Its Classification, Prediction, and Clustering Optimization in Aceh, Indonesia*," Informatics, vol. 11, no. 4, p. 89, 2024.

[20] N. Nurdin, Fajriana, Rini Meiyanti, Adelia, and Maya Maulita, "Clustering and Mapping of Agricultural Production Based on Geographic Information System Using K-Medoids Algorithm", *JAIT*, vol. 5, pp. 116–124, Feb. 2025.

[21] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "*Cluster validity indices for automatic clustering: A comprehensive review*," Heliyon, vol. 11, no. 2, 2025.

[22] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, "*Analysis of microarray data using Z score transformation*," Journal of Molecular Diagnostics, vol. 5, no. 2, pp. 73–81, 2003.

[23] R. K. Dinata, R. T. Adek, N. Hasdyna, and S. Retno, "*K-nearest neighbor classifier optimization using purity,*" AIP Conference Proceedings, vol. 2431, no. 1, AIP Publishing, Aug. 2023.

[24] F. M. Hasan, T. F. Hussein, H. D. Saleem, and O. S. Qasim, "*Enhanced unsupervised feature selection method using crow search algorithm and Calinski-Harabasz*," International Journal of Computational Methods and Experimental Measurements, vol. 12, no. 2, pp. 185–190, 2024.

[25] P. Palli, S. Mishra, and P. S. Rao, "*Inferring compound similarity: a clustering approach in drug discovery*," in 2024 1st International Conference on Cognitive, Green and Ubiquitous Computing (IC-CGU), IEEE, pp. 1–6, Mar. 2024.