International Journal of Management Science and Information Technology
IJMSIT

E-ISSN: 2774-5694 P-ISSN: 2776-7388 OPEN ACCESS

Volume 6 (1), January-June 2026, 297-304
DOI: https://doi.org/10.35870/ijmsit.v6i1.6648

Optimizing K-means Clustering with Seed Initialization for Osteoporosis Diagnosis Based on Family History

Adiyah Mahiruna [1*], Ngatimin [2], Rachmat Destriana [3]
[1*,2] Software Engineering Study Program, Faculty of Science and Technology, Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang, Semarang City, Central Java Province, Indonesia
[3] Informatics Engineering Study Program, Faculty of Engineering, Universitas Muhammadiyah Tangerang, Tangerang City, Banten Province, Indonesia

Email: mahirunaadiyah@gmail.com [1*], ngatimin@itesa.ac.id [2], rahmat.destriana@ft-umt.ac.id [3]

Article history:
Received February 21, 2026
Revised March 27, 2026
Accepted April 1, 2026

Abstract
World Osteoporosis Day (WOD) is celebrated on October 20 every year, to raise global awareness about the prevention, diagnosis, and treatment of osteoporosis. Urgency in Indonesia, the number of elderly people is projected to reach 71 million people in 2050, which will have an impact on increasing cases of osteoporosis. Therefore, the recommendations based on scientific evidence in this study aim to assist practitioners in preventing osteoporosis in adults and children. This study proposes a method of Improving K-Means Performance through Seeds. The performance of the K-Means clustering algorithm is highly dependent on the random selection of initial centroids, which can lead to unstable clusters, suboptimal local solutions, and increased iterations, particularly in medical datasets such as osteoporosis diagnosis based on family history. Therefore, there is a need for an optimized centroid initialization strategy that can improve clustering accuracy and stability without increasing computational complexity. The dataset used is the osteoporosis dataset as a testing dataset that can be accessed publicly Osteoporosis dataset. The novelty of this study lies in the introduction of Modified Average (MA) approach for centroid initialization, which eliminates random seed dependency and improves clustering stability without increasing computational complexity. From the results of nine experiments with the benchmarking dataset, it can be seen that the method proposed in this study indicates that practically the Proposed method has a tendency to perform better in Rand Index measurement compare to k-means in random seeds.

Keywords:
K-means; Seeds; Clustering; Osteoporosis; Rand index.

1. INTRODUCTION

Osteoporosis is known as brittle bones. Initially, osteoporosis only affected the health of the elderly, especially those who are elderly. Advances in knowledge today also show that osteoporosis occurs in children, either as a precursor to osteoporosis or even as a predictor of osteoporosis in the elderly population (Kemenkes RI, 2022). In Indonesia, clustering has not been carried out for patients diagnosed with osteoporosis based on family history. One of the factor's causing osteoporosis is family history, if a family member has osteoporosis, then the risk of that person experiencing it becomes greater. Clustering plays an important role in analyzing and grouping osteoporosis sufferers based on risk factors such as family history or genetic factors (Laurenso et al., 2024; Sitinjak et al., 2022). By using clustering techniques, individuals with similar characteristics can be grouped to understand the pattern of association between hereditary factors and the development of osteoporosis (. et al., 2024; Mahmuda et al., 2023). Clustering methods for family health history are carried out to prevent the increasing number of osteoporosis sufferers.

The K-Means method is a clustering algorithm (Daulay & Wandri, 2025) that partitions data by performing an iterative process in forming data groups (Faran & Aldisa, 2024; Indra et al., 2024) through a series of iterative partitions to reduce the average distance between each data and the corresponding cluster

center. According to the performance assessment of the K-means and Fuzzy C-means segmentation techniques in conjunction with 3 ML, the osteoporosis detection method demonstrating the highest diagnostic performance is K-means segmentation paired with a multilayer perceptron classifier, achieving accuracies of 90.48%, 90.90%, and 90.00% for specificity and sensitivity, respectively (Widyaningrum et al., 2023). However, the performance of K-means is highly dependent on the selection of the initial center (centroid) which is determined randomly (Chen et al., 2020; Erisoglu et al., 2011; Lu & Braunstein, 2014), and depends on the determination of the number of clusters (Naldi & Campello, 2014). Suboptimal initial center selection can lead to convergence to less accurate local solutions (Goyal & Kumar, 2014; Tsapanos et al., 2015), produce unstable clusters, and increase the number of iterations required to reach the final result (Farissa et al., 2021; Maulani et al., 2025). Therefore, an optimization method is needed in determining the initial K-Means centers so that the clustering process is more effective and produces more accurate data groups (Celebi, 2015; Celebi & Kingravi, 2012). In research by Ahmad Ilham [20] the findings demonstrated that the suggested approach produces great SSE values, particularly for k=4, which has the lowest SSE value as opposed to k=3, itt has been demonstrated that applying DT to enhance Goyal and Kumar's approach (Goyal & Kumar, 2014) to the initial centroid improves k-means performance. In research by S.A. Sajidha (Sajidha et al., 2018) proposed initial seed artifacts for the K-modes technique is the primary goal of the algorithm the researchers present in their paper. In order to select the seed artifacts from distinct clusters and dense places.

Based on these problems that seeds of k-means is important, many researchers have conducted research on determining k-means seeds. Then this study focuses on how to determine the optimal initial center to improve the performance of the K-Means algorithm in clustering, especially in the application of osteoporosis diagnosis based on family history factors. The method used to determine the initial center of the K-means algorithm is the modified average (MA).

2. RESEARCH METHOD

In this study, a public dataset sourced from Kaggle Osteoporosis dataset and the University of California Irvine (UCI) will be used.

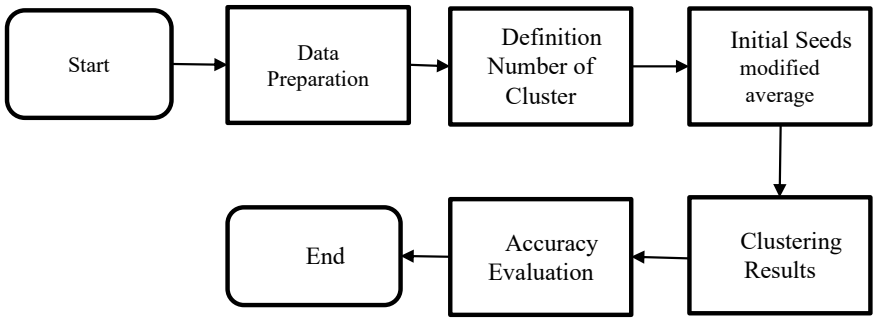

Figure 1. The Flow of Proposed Method

The processing of proposed method using the R studio application, the k-means that konventional and the proposed Seeds for k-means is processing with the same datasets.

2.1. K-mean Method

The steps in the K-means method generally include the following random seeds: 1) Prepare the datasets, 2) Determine the number of clusters, 3) Random Seeds, 4) Calculate the distance to the Seeds, Euclidean distance, 5) Grouping based on minimum distance, 6) No Object Moved Group, 7) The Clusters is created.

The k-means method is random seeds for initial first centroid, the proposed method using modified average (MA) to initial the first centroid.

2.2. Proposed Method

The steps of proposed method: 1) Prepare the datasets, 2) Determine the number of clusters, 3) Modified average (MA) Seeds, 4) Calculate the distance to the Seeds, Euclidean distance, 5) Grouping based on minimum distance, 6) No Object Moved Group, 7) The Clusters is created.

The proposed method and the conventional k-means is different in the 3rd step, we can see in the two steps above that the difference between the proposed method and the conventional k-means method is in the third step. The Existing study using average in every part of dataset that partising base on the number of k, this proposed method using global average without partising the dataset, the proposed method in this paper using global average then divide the result of average with the number of k. Figure 2 explain the modified average (MA) method that proposed in this study.

```
INPUT:
   D = {d1, d2, d3, …, dN}
   K
   Label [1,..N]
OUTPUT
   count [1..K]
   mean [1..K]
ALGORITHM
1. Initialization
   FOR k ← 1 TO K DO
     count[k] ← 0
     Sum[k] ← 0
   END FOR
2. Computer the number of data points and total value for each cluster
FOR i ← 1 TO N DO
   K ← label [i]
   count [k] ← count[k] + 1
   sum[k] ← sum[k] + D[i]
END FOR
3. Compute the mean of each cluster
FOR k □ 1 TO k DO
   If count[k] != 0 THEN
   mean [k] ← sum[k] / count[k]
   Else
   mean[k] ← 0
   End if
END FOR
4. Return count and mean
```

Figure 2. Proposed Method

From Figure 2. We can know the different about the step of the proposed method, we will explain the different our proposed method with existing method.

2.3. Basic K-Means Equation

(1) Random seeds
(2) Euclidean Distance

$$d(x_i, \mu_j) = \sqrt{\sum_{i=1}^{m} (x_{il} - \mu_{jl})^2} \quad (1)$$

For:
m = Number of Attribute
Xil = value of l-th attribute from i-th data
μjl = value of l-th attribute from j-th data

(3) Update Centroid

$$(\mu_i) = \frac{1}{C_i} \sum_{x_j \in C_i} x_j \quad (2)$$

For:
Ci = number of members of cluster i

2.4. Proposed Method Equation (Modified Average)

(1) Modified Average for seeds

$$MA = \frac{1}{n} \sum_{i=1}^{n} x_i \quad (3)$$

For:
n = Numbers of data
Xi = i-th data

$$MA_1 = \frac{1}{k}\left(\frac{1}{n}\sum_{i=1}^{n} x_{il}\right) \quad (4)$$

For:
k = number of Cluster
l = i-th attribute

(2) Euclidean Distance

$$d(x_i, \mu_j) = \sqrt{\sum_{i=1}^{m}(x_{il} - \mu_{jl})^2} \quad (5)$$

For:
m = Number of Attribute
Xil = value of l-th attribute from i-th data
μjl = value of l-th attribute from j-th data

(3) Update Centroid

$$(\mu_i) = \frac{1}{C_i} \sum_{x_j \in C_i} x_j \quad (6)$$

For:
Ci = number of members of cluster i

2.5. Rand Index

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \quad (7)$$

For:
TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

2.6. Confidence Interval (CI)

$$CI = [D_{min}, D_{max}] \quad (7)$$

For:
CI = Confidence Interval
TP = Minimal Difference
TN = Maximal Difference

3. RESULTS AND DISCUSSION

In this study, the method used to determine the initial cluster center is by applying the modified average (MA) method to determine the initial cluster center. The performance of the tested method was measured using the Rand Index. The public dataset sourced from Kaggle Osteoporosis dataset and the University of California Irvine (UCI): BreastTissue and Immunotherapy. The proposed method compares to k-means random seeds.

Table 1. Dataset Used in This Study

No	Dataset Name	Data Amount	Number of Attributes	Number of Classes
1	Osteoporosis	1958	10	2
2	Breast Tissue	106	9	4
3	Immunotherapy	90	7	2

In this Study the method will be process in R Studio application and using public dataset to do experience the proposed and analyze the results.

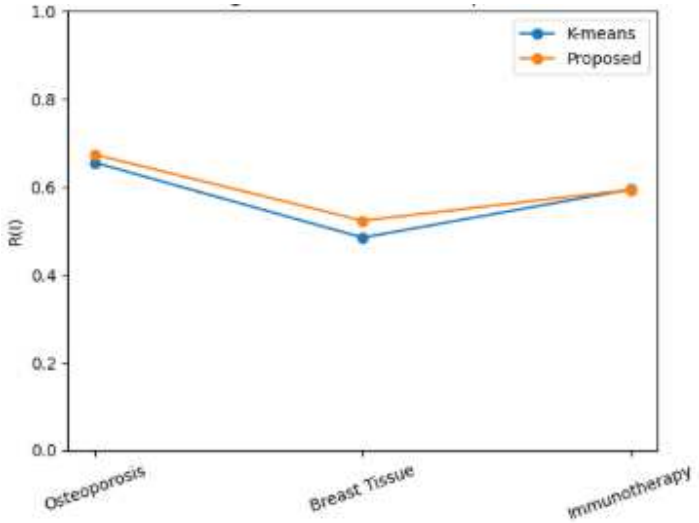

Figure 3. R(I) Diagram: K-means and Proposed Method

From Figure 3. We can see that the diagram shows how the interpretations are Osteoporosis: Proposed slightly higher than K-means; Breast Tissue: Proposed to be higher by the most obvious difference; Immunotherapy: Both values are almost the same (Proposed slightly lower). The Rand Index value ranges from 0 to 1. The greater the Rand Index value, the more similar the attribute data between members in one cluster.

Family history is a significant risk factor, as individuals with relatives affected by osteoporosis are more likely to develop the condition themselves. Identifying groups of individuals with similar characteristics can aid in early detection and preventive interventions. Clustering techniques, particularly K-Means, are widely used for this purpose due to their simplicity and efficiency in partitioning data based on centroid distances. However, the performance of K-Means is highly dependent on the initial selection of centroids and the predetermined number of clusters. Random initialization often leads to unstable clusters, suboptimal solutions, and increased iterations. Several studies have explored deterministic or optimized initialization methods to improve clustering outcomes, but there remains a need for approaches that reduce randomness while maintaining computational efficiency.

This study proposes a Modified Average (MA) method for initializing K-Means centroids to address these challenges. Unlike conventional K-Means, which selects initial centroids randomly, the MA method computes the global average of all attributes and divides this average into kkk initial centroids. This approach reduces dependency on random seeds, improves clustering stability, and does not increase computational complexity. Experiments were conducted using publicly available datasets, including an osteoporosis dataset with 1,958 samples and 10 attributes, the Breast Tissue dataset with 106 samples and 9 attributes, and the Immunotherapy dataset with 90 samples and 7 attributes. Data preprocessing, including normalization and handling missing values, ensured consistency across experiments. The performance of the proposed method was evaluated using the Rand Index, which measures the similarity between predicted clusters and actual classifications. Statistical analysis, including confidence intervals calculated via the Wilcoxon signed-rank test, was used to assess differences between the proposed method and conventional K-Means.

Table 2. Results R(I) with Number of Cluster (K) = 3

No	Dataset Name	Number of Class	R(I) K-means	R(I) Proposed	Difference (D)
1	Osteoporosis	2	0.65409	0.6723	0.01821
2	Breast Tissue	4	0.4832	0.5219	0.0387
3	Immunotherapy	2	0.5930	0.5922	-0.0008

The proposed method has greater R(I), the proposed method has a higher rand index value than conventional k-means on dataset Osteoporosis and Breast Tissue, in dataset immunotherapy the proposed method has 0.001 lower.

In this study we use Confidence Interval (CI) is used for statistical analysis. Confidence Interval Performance Difference is in Table 3.

Table 3. Confidence Interval Performance Difference (D)

No	Dataset Name	Number of Class	R(I) K-means	R(I) Proposed	Difference (D)
1	Osteoporosis	2	0.65409	0.6723	0.01821
2	Breast Tissue	4	0.4832	0.5219	0.0387
3	Immunotherapy	2	0.5930	0.5922	-0.0008

Wilcoxon signed-rank distribution used in this study to compare the means of two paired or dependent groups by analyzing the difference in ranks from non-normally distributed data. For small n (n = 3), the non-parametric Confidence Interval (CI) is calculated based on the Wilcoxon rank distribution. With n = 3 and α = 0.05, the 95% confidence interval for the median difference is between: CI = [Dmin, Dmax] = [-0.0008, 0.0387]. The performance difference between the Proposed and K-means methods is not statistically significant at the 95% confidence level, but 2 of 3 datasets show an increase and the upper limit of the CI indicates a potential increase of up to 0.0387. This indicates that practically the Proposed method has a tendency to perform better, although statistically it is not yet strong enough.

Results indicate that the MA-based method consistently improves clustering performance compared to conventional K-Means. On the Osteoporosis dataset, the Rand Index increased from 0.6541 to 0.6723, while on the Breast Tissue dataset, it increased from 0.4832 to 0.5219. The Immunotherapy dataset showed negligible difference, with the Rand Index slightly decreasing from 0.5930 to 0.5922. Although the statistical significance of these improvements is limited due to the small number of datasets, the practical improvement demonstrates the effectiveness of the MA initialization. The proposed method provides more stable clusters, reduces variability due to random seed selection, and performs consistently across different datasets. Future work should involve larger and more diverse datasets and include additional clustering metrics such as Silhouette Score, Adjusted Rand Index (ARI), Davies-Bouldin Index, and Normalized Mutual Information (NMI) to provide a comprehensive evaluation. Integrating this approach with predictive modeling could also enhance early diagnosis and preventive strategies for osteoporosis.

4. CONCLUSION

Base from the results of the experiments, the proposed method performed better than K-means on two of the three datasets, namely Osteoporosis and Breast Tissue. On the Immunotherapy dataset, the performance of both methods was relatively equivalent. The 95% confidence interval for the median performance difference is in the range [-0.0008, 0.0387], which includes the value of zero. This confirms that the performance improvement of the Proposed method is not yet statistically significant. The insignificance of the statistical test results was greatly influenced by the limited number of datasets (n = 3), so the statistical power of the test was relatively low. Based on the results of research and analysis that has been carried out so the future work are to increase the power of statistical tests and obtain more representative conclusions, it is recommended to use more datasets with diverse characteristics, In addition to R(I), it is recommended to use other metrics such as Silhouette Score, Adjusted Rand Index (ARI), Davies-Bouldin Index, or Normalized Mutual Information (NMI) to obtain a more comprehensive evaluation.

The Modified Average K-Means method offers a simple yet effective enhancement over conventional K-Means by providing deterministic initial centroids that improve cluster stability and performance. While improvements are practically meaningful, further studies with more datasets and complementary evaluation metrics are necessary to establish statistical significance and broader applicability. This approach can serve as a valuable tool for medical data analysis, particularly in identifying high-risk populations for osteoporosis based on family history.

ACKNOWLEDGEMENTS

The author would like to thank the Directorate of Research, Technology, and Community Service (DPPM) for the financial support that has made this research possible.

REFERENCES

Celebi, M. E. (2015). Partitional clustering algorithms. Partitional Clustering Algorithms, September, 1–415. https://doi.org/10.1007/978-3-319-09259-1

Celebi, M. E., & Kingravi, H. A. (2012). Deterministic initialization of the K-means algorithm using hierarchical clustering. International Journal of Pattern Recognition and Artificial Intelligence, 26(7). https://doi.org/10.1142/S0218001412500188

Chen, J., Qi, X., Chen, L., Chen, F., & Cheng, G. (2020). Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection. Knowledge-Based Systems, 203, 106167. https://doi.org/10.1016/j.knosys.2020.106167

Daulay, S., & Wandri, R. (2025). Integrating K-Means Clustering and K-Nearest Neighbor Classification for Effective Scholarship Recipient Selection. Sistemasi: Jurnal Sistem Informasi, 14(1), 235-248. https://doi.org/10.32520/stmsi.v14i1.4818

Erisoglu, M., Calis, N., & Sakallioglu, S. (2011). A new algorithm for initial cluster centers in k-means algorithm. Pattern Recognition Letters, 32(14), 1701–1705. https://doi.org/10.1016/j.patrec.2011.07.011

Faran, J., & Aldisa, R. T. (2024). Perbandingan Algoritma K-Means dan K-Medoids Dalam Pengelompokan Kelas Untuk Mahasiswa Baru Program Magister. Journal of Information System Research (JOSH), 5(2), 509–519. https://doi.org/10.47065/josh.v5i2.4753

Farissa, R. A., Mayasari, R., & Umaidah, Y. (2021). Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient di Puskesmas Karangsambung. Journal of Applied Informatics and Computing, 5(2), 109–116. https://doi.org/10.30871/jaic.v5i1.3237

Goyal, M., & Kumar, S. (2014). Improving the Initial Centroids of k-means Clustering Algorithm to Generalize its Applicability. Journal of The Institution of Engineers (India): Series B, 95(4), 345–350. https://doi.org/10.1007/s40031-014-0106-z

Indra, I. I., Rizki, U., Jakak, P. M., Prayogi, M. B., & Rahman, M. (2024). Penerapan Metode K-Means Clustering Dalam Pengembangan Strategi Promosi Berbasis Data Penerimaan Mahasiswa Baru (Studi Kasus: Universitas Nurul Huda). Jurnal Nasional Ilmu Komputer, 5(1), 25–43. https://doi.org/10.47747/jurnalnik.v5i1.1656

Kemenkes RI. (2022). Kepmenkes RI no HK.01.07/MENKES/1928/2022 Tentang Pedoman Nasional Pelayanan Kedokteran Tata Laksana Stunting. 1–52.

Laurenso, J., Jiustian, D., Fernando, F., Suhandi, V., & Rochadiani, T. H. (2024). Implementation of K-Means, Hierarchical, and BIRCH Clustering Algorithms to Determine Marketing Targets for Vape Sales in Indonesia. Journal of Applied Informatics and Computing, 8(1), 62–70. https://doi.org/10.30871/jaic.v8i1.4871

Lu, S., & Braunstein, S. L. (2014). Quantum decision tree classifier. Quantum Information Processing, 13(3), 757–770. https://doi.org/10.1007/s11128-013-0687-5

Mahmuda, F., Sitorus, M. A. R., Widyastuti, H., & Kurniawan, D. E. (2018). Clustering Profil Pengunjung Perpustakaan Menggunakan Algoritma K-Means: (Studi Kasus Perpustakaan BP Batam). Journal of Applied Informatics and Computing, 1(1), 14–21. https://doi.org/10.30871/jaic.v1i1.476

Maulani, V. R., Barata, M. A., & Yuwita, P. E. (2025). A Improving House Price Clustering Results with K-means through the Implementation of One-hot Encoding Pre-processing Technique. Journal of Applied Informatics and Computing, 9(3), 741–748. https://doi.org/10.30871/jaic.v9i3.9481

Naldi, M. C., & Campello, R. J. G. B. (2014). Evolutionary k-means for distributed data sets. Neurocomputing, 127, 30–42. https://doi.org/10.1016/j.neucom.2013.05.046

R., Lapatta, N. T., Ardiansyah, R., . W., & Angreni, D. S. (2024). Donor Segmentation Analysis Using the RFM Model and K-Means Clustering to Optimize Fundraising Strategies. Journal of Applied Informatics and Computing, 8(2), 341–349. https://doi.org/10.30871/jaic.v8i2.8464

Sajidha, S. A., Chodnekar, S. P., & Desikan, K. (2018). Initial seed selection for K-modes clustering – A distance and density based approach. Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2018.04.013

Sitinjak, D. K., Pangestu, B. A., & Sari, B. N. (2022). Clustering Tenaga Kesehatan Berdasarkan Kecamatan di Kabupaten Karawang Menggunakan Algoritma K-Means. Journal of Applied Informatics and Computing, 6(1), 47–54. https://doi.org/10.30871/jaic.v6i1.3855

Tsapanos, N., Tefas, A., Nikolaidis, N., & Pitas, I. (2015). A distributed framework for trimmed Kernel k - Means clustering. Pattern Recognition, 48(8), 2685–2698. https://doi.org/10.1016/j.patcog.2015.02.020

Widyaningrum, R., Sela, E. I., Pulungan, R., & Septiarini, A. (2023). Automatic Segmentation of Periapical Radiograph Using Color Histogram and Machine Learning for Osteoporosis Detection. International Journal of Dentistry, 2023. https://doi.org/10.1155/2023/6662911