The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 5 | Issue No. 8 | August 2025 Impact Factor: 6.53

Clustering Techniques for Heart Attack Prediction: A Silhouette Score-Based Evaluation

Paras Negi (parasnet072@gmail.com), Corresponding Author
Research Scholar, Soban Singh Jeena University, Almora Campus, Almora, Uttarakhand, India
Manoj Kumar Bisht (manojssj2020@gmail.com)
Assistant Professor, Soban Singh Jeena University, Almora Campus, Almora, Uttarakhand, India

 Crossref/DOI: https://doi.org/10.55454/rcsas.5.08.2025.002

Abstract: *Heart attacks, also known as myocardial infarctions, remain a major contributor to global mortality, necessitating early detection of individuals at risk. Traditional diagnostic methods often fail to uncover complex patterns in medical data. This study explores the use of unsupervised machine learning techniques (K-Means, K-Medoids, and Agglomerative Clustering) to segment patients based on cardiovascular risk factors and support heart attack prediction. A publicly available heart attack dataset was preprocessed using normalization and label encoding. The clustering outcomes were evaluated using the Silhouette Score. Among the three techniques, K-Medoids achieved the highest Silhouette Score, indicating more coherent and meaningful patient groupings.*

Keywords: Clustering Techniques, Heart Attack Prediction, Machine Learning, Silhouette Score.

Article History: **Received: 04 August- 2025;** **Accepted: 15 August- 2025;** **Published/Available Online: 30 August- 2025**

I. Introduction

Cardiovascular diseases, particularly heart attacks, remain the leading cause of death worldwide, accounting for approximately 17.9 million deaths annually, according to the World Health Organization [1]. Early detection and accurate risk prediction are essential for reducing mortality rates and improving patient outcomes. However, conventional diagnostic approaches often rely on rule-based systems and clinical expertise, which may fall short in capturing the complex, nonlinear relationships present in patient data. In recent years, machine learning (ML) has emerged as a transformative tool in healthcare analytics, offering enhanced capabilities for early diagnosis, prognosis, and risk stratification in cardiovascular medicine [2]. Among ML approaches, unsupervised learning, particularly clustering, has shown significant promise in identifying latent patterns in unlabeled clinical data [3]. Clustering enables the segmentation of patients into distinct risk categories based on shared clinical features, thereby facilitating more personalized and proactive medical interventions [4].

This study investigates the use of three widely adopted clustering techniques **(**K-Means**,** K-Medoids**,** and Agglomerative Clustering**)** to group patients according to cardiovascular risk factors [5]. Unlike many existing studies that emphasize supervised classification, this study highlights the value of unsupervised learning in revealing natural structures within medical data that may otherwise go unnoticed.

The experimental analysis is conducted using a publicly available heart disease dataset from Kaggle [6], originally sourced from the UCI Machine Learning Repository. The Silhouette Score, an internal validation metric that evaluates both intra-cluster cohesion and inter-cluster separation, has been employed to evaluate clustering performance.

II. Related Work

Given the increasing availability of healthcare data, machine learning techniques, particularly clustering, have become essential tools for the early detection and risk stratification of heart attacks. Clustering has been widely applied in the medical domain to uncover hidden patterns in patient data without the need for predefined class labels [7]. In the context of heart attack prediction, clustering enables the grouping of patients based on similarities in clinical features, which are often indicative of varying cardiovascular risk levels [8]. K-means is a widely used clustering technique known for its simplicity, interpretability, and computational efficiency, and has been extensively applied to categorize patients based on heart attack risk levels [3]. Other techniques, such as Hierarchical Clustering and DBSCAN, have also gained attention. DBSCAN is particularly advantageous for detecting outliers and identifying clusters of arbitrary shapes, which makes it suitable for handling complex and diverse medical datasets [9]. While the selection of an appropriate clustering technique is important, evaluating the quality of clusters is equally crucial. For this, a

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 5 | Issue No. 8 | August 2025 Impact Factor: 6.53

frequently used internal validation metric is the silhouette score. It measures how similar an object is to its cluster compared to other clusters, with values closer to +1 indicating better-defined clusters [10].

Silhouette analysis is also used to determine the optimal number of clusters, particularly in health datasets such as those involving heart disease [11]. Several studies have applied the silhouette score to evaluate clustering performance in the context of cardiovascular disease. Sharma and Patel [12], for example, analyzed the Cleveland Heart Disease Dataset using K-means, DBSCAN, and Agglomerative Clustering. Their study found that DBSCAN yielded the best silhouette score, attributed to its flexibility in forming non-spherical clusters and identifying noisy data points. Recent research has also explored hybrid models that combine clustering with classification techniques to improve prediction accuracy. In such approaches, clustering is used in preprocessing or feature engineering, while supervised learning algorithms such as Random Forest perform the final classification. The silhouette score plays a pivotal role in validating the clustering step before moving to classification [13].

In addition, Rajesh and Reena [14] implemented K-means clustering for recognizing patient patterns, although they did not incorporate internal validation metrics such as the silhouette score—highlighting a gap in comprehensive evaluation. Chaurasia and Pal [15] further enhanced clustering methods by integrating feature selection techniques, leading to improved diagnostic performance.

III. Methodology

This section outlines the methodology used for predicting heart attack risk using unsupervised clustering techniques applied to a publicly available dataset from Kaggle [6]. The process consisted of three main phases: data preprocessing, implementation of clustering techniques, and evaluation using the Silhouette Score.

A. Dataset Description

The dataset used in this study was obtained from Kaggle [6] and originally sourced from the UCI Machine Learning Repository. It consists of 1,025 records of patient data, including clinical features such as age, sex, resting blood pressure, cholesterol level, fasting blood sugar, maximum heart rate achieved, and exercise-induced angina. The target variable indicating the presence or absence of heart attack was excluded during clustering, as unsupervised learning does not utilize labeled outcomes.

B. Data Preprocessing

Before applying clustering techniques, the dataset underwent the following preprocessing steps:

Handling Missing Values:

The dataset contained no null values. However, categorical variables were label-encoded where necessary to convert them into numerical form suitable for clustering [3].

Feature Scaling:

To ensure uniform contribution of features during clustering, multiple normalization techniques were evaluated, including Min-Max Scaler, Standard Scaler, and Robust Scaler [3].Min-Max Scaling transforms numerical attributes into a fixed range, typically [0, 1], by rescaling each feature based on its minimum and maximum values. Standard Scaling (Z-score normalization) standardizes features by removing the mean and scaling to unit variance. Robust Scaling uses the median and interquartile range, making it less sensitive to outliers.

Dimensionality Reduction Technique:

For visualization purposes, Principal Component Analysis (PCA) was applied to reduce the dataset's dimensionality to two principal components, facilitating cluster visualization [3].

C. Clustering Techniques

Three clustering techniques were employed to group patients based on clinical similarity:

- **K-Means:** K-Means is a widely used centroid-based algorithm that partitions data into k clusters by minimizing the sum of squared distances between data points and their respective cluster centroids.

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 5 | Issue No. 8 | August 2025 Impact Factor: 6.53

While efficient and scalable, it is sensitive to outliers and assumes clusters are spherical and of similar size [3].

- **K-Medoids:** K-Medoids, proposed by Kaufman and Rousseeuw [5], provide a more robust alternative to K-Means. Unlike K-Means, it selects actual data points (medoids) as cluster centers and minimizes the sum of pairwise dissimilarities between each point and its nearest medoid. Consequently, it handles noise and outliers more effectively.
- **Agglomerative Clustering:** Agglomerative Clustering is a type of hierarchical clustering that builds nested clusters by successively merging the most similar pairs of data points based on a defined linkage criterion. It begins by treating each data point as an individual cluster and iteratively merges them until the desired number of clusters is reached [7].

D. Evaluation Metric

The Silhouette Score, used to evaluate the quality of clustering, quantifies both the compactness and separation of clusters. It provides an internal validation measure that does not require labeled data, making it well-suited for unsupervised learning tasks [3].The Silhouette Score S for a data point is defined as:

$$S = (b - a) / \max(a, b)$$

Here, a and b represent the average distances to points inside the same cluster and the nearest neighboring cluster, respectively. Clustering quality is reflected by the Silhouette Score, which ranges from -1 to +1. Scores close to +1 indicate that the data points are well-clustered and separated from neighboring clusters. Values near 0 suggest that the clusters are overlapping, and scores approaching –1 imply that the data points may have been incorrectly assigned, indicating poor clustering or misclassification. The clustering technique that achieved the highest average Silhouette Score across all data points was considered the most effective for this dataset.

IV. Experimental Results

A. Clustering Performance

Figure 1 presents the performance of various clustering techniques based on the Silhouette Score. To enhance the robustness and generalizability of the results, each clustering technique was tested under multiple preprocessing conditions, including three feature scaling methods(Standard Scaler, Min-Max Scaler, and Robust Scaler) as well as different initialization strategies for K-Medoids (k-medoids++, random, and heuristic initializations).

B. Effect of Preprocessing

The choice of feature scaling method significantly influenced clustering performance. Min-Max Scaler consistently yielded superior results across all clustering techniques. Its ability to normalize features into a uniform range of [0, 1] enhanced the reliability of Euclidean distance-based computations, thereby improving clustering accuracy. Robust Scaler delivered moderate results in some configurations but failed to demonstrate consistent improvements. Its performance was inferior in high-performing scenarios and did not outperform Min-Max scaling in any case. Standard Scaler, although commonly applied in general ML workflows, was consistently outperformed by Min-Max Scaler across all tested techniques-scaler combinations.

C. Interpretation of Silhouette Scores

Top-performing configurations included K-Medoids with Min-Max Scaler and k-medoids++ initialization, as well as K-Means with Min-Max Scaler. These combinations achieved average Silhouette Scores between 0.26 and 0.27, suggesting moderately strong clustering structures. Poorly performing configurations, such as K-Medoids combined with Standard Scaler and heuristic initialization, produced negative Silhouette Scores, indicating overlapping clusters, poor separation, and a high probability of misclassification.

The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com
ISSN: 2583-1380 Vol. 5 | Issue No. 8 | August 2025 Impact Factor: 6.53

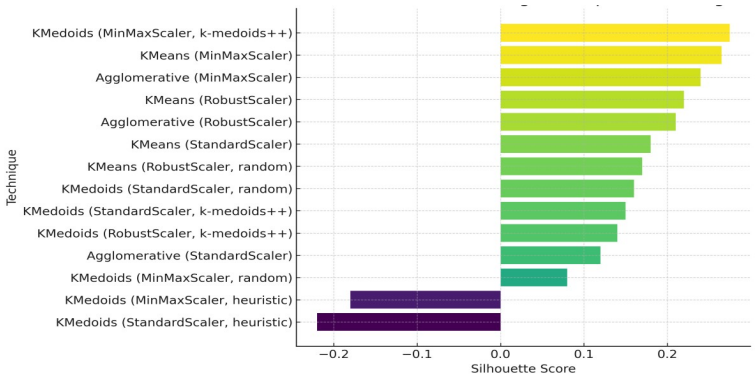

Figure 1: Silhouette Scores of Clustering Techniques

The dataset was reduced to two dimensions using Principal Component Analysis in order to visually evaluate clustering performance. This dimensionality reduction technique allows the high-dimensional data to be projected onto a 2D plane, enabling graphical representation of cluster boundaries and data distribution. Figure 2 illustrates the clustering outcomes for the different techniques applied to the heart attack dataset. The plots reveal how each technique grouped the data points, highlighting differences in cluster density, separation, and overlap. Among the configurations, K-Medoids with Min-Max scaling exhibited the most distinct and compact clusters, aligning with its superior Silhouette Score and suggesting its effectiveness in capturing meaningful patient subgroups.

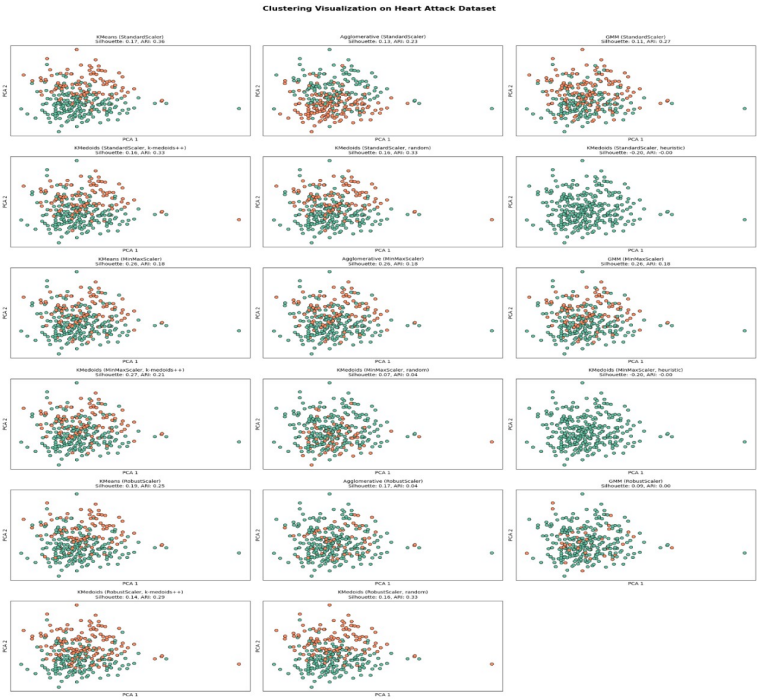

Figure 2: Clustering Visualization on the Heart Attack Dataset

V. Conclusion and Future Work

A. Conclusion

This study evaluated the effectiveness of K-Means, K-Medoids, and Agglomerative Clustering for segmenting patients based on cardiovascular risk factors. Through rigorous experimentation on a publicly available heart attack dataset, clustering performance was evaluated using the Silhouette Score. Among these techniques, K-Medoids, particularly in combination with Min-Max scaling and k-medoids++ initialization, achieved the highest Silhouette Score. This indicates its superior ability to form well-defined and distinct clusters compared to K-Means and Agglomerative Clustering. These results demonstrate that unsupervised

learning can reveal latent subgroups within patient populations, offering promising insights for risk stratification, early intervention, and personalized healthcare planning, even in the absence of labeled data.

B. Future Work

While the findings are promising, several directions remain for future exploration. One potential direction is the use of hybrid and ensemble clustering approaches, which combine multiple clustering techniques to improve stability and uncover complementary patterns in complex clinical data. Semi-supervised learning could also be explored, where incorporating a small amount of labeled data may help refine cluster boundaries and improve clinical interpretability. Additionally, integrating deep learning approaches (such as autoencoders or variational autoencoders) could enable the discovery of more complex, nonlinear relationships in high-dimensional datasets. Expanding the analysis to larger and multi-source datasets may further improve the generalizability and clinical relevance of the clustering framework. Real-world validation is essential to assess whether the identified clusters align with clinical diagnoses, prognostic outcomes, or treatment responses in actual healthcare settings.

References

[1] World Health Organization, “Cardiovascular diseases (CVDs),” Jun. 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] M. Mohan, C. Thirumalai, and G. Srivastava, “Effective heart disease prediction using hybrid machine learning techniques,” IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[3] J. Han, M. Kamber, and J. Pei, “Data Mining: Concepts and Techniques”, 3rd ed. Amsterdam, Netherlands: Morgan Kaufmann, 2011.

[4] S. Arora and S. Sharma, “ Comparative analysis of clustering techniques for heart disease prediction", in Proc. 3rd Int. Conf. Comput. Commun. Autom. (ICCCA), Greater Noida, India, 2017, pp. 1–6, doi: 10.1109/CCAA.2017.8229926.

[5] L. Kaufman and P. J. Rousseeuw, “Finding Groups in Data: An Introduction to Cluster Analysis”. Hoboken, NJ, USA: Wiley, 2009.

[6] Kaggle, “Heart Disease UCI Dataset.” [Online]. Available: https://www.kaggle.com/datasets/ronitf/heart-disease-uci

[7] A. K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, 2010.

[8] M. Gudadhe, K. Wankhade, and S. Dongre, “Decision support system for heart disease based on support vector machine and artificial neural network”, in Proc. Int. Conf. Comput. Commun. Technol. (ICCCT), 2010, pp. 741–745.

[9] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise”, in Proc. 2nd Int. Conf. Knowl. Discov. Data Min. (KDD), 1996, pp. 226–231.

[10] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”, J. Comput. Appl. Math., vol. 20, pp. 53–65, 1987.

[11] N. Garg, N. Sharma, and A. Kumari, “Comparative study of clustering algorithms using silhouette score”, Int. J. Eng. Res. Technol., vol. 10, no. 5, pp. 234–238, 2021.

[12] N. Sharma and R. Patel, “Performance analysis of clustering algorithms on heart disease data”, Int. J. Comput. Sci. Eng., vol. 6, no. 6, pp. 1076–1081, 2018.

[13] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction”, BMC Med. Inform. Decis. Mak., vol. 19, no. 1, pp. 1–16, 2019, doi: 10.1186/s12911-019-1004-8.

[14] K. Rajesh and G. S. Reena, “Analysis of heart disease dataset using neural network approach,” Int. J. Comput. Sci. Inf. Technol., vol. 2, no. 3, pp. 22–28, 2011.

[15] V. Chaurasia and S. Pal, “A novel approach for heart disease prediction using data mining techniques,” Int. J. Eng. Res. Gen. Sci., vol. 2, no. 1, pp. 315–319, 2014.

Conflict of Interest: The authors declare “No conflict of Interest”.