Evaluating Clustering Methods for Heart Disease Analysis

Y.K.K.M.K. Gayathri and N.A.D.N. Napagoda

Abstract— Heart disease is a major worldwide health concern, demanding the creation of effective diagnostic and prediction techniques. Clustering methods, which have the ability to identify hidden patterns within patient populations, have emerged as valuable tools for studying medical data. In this study, the evaluation of two prominent clustering methods, Hierarchical clustering and K-means clustering, in the context of heart disease classification, is the central focus. A comprehensive dataset specifically curated for heart disease research is utilized in this study, and it delves into the performance of these clustering techniques regarding their capacity to identify distinct patient subgroups. The dataset undergoes rigorous preprocessing steps, including feature standardization, to ensure the reliability of the analysis. The results indicate that K-means clustering surpasses Hierarchical clustering in terms of clustering accuracy, achieving a remarkable 81% accuracy rate compared to the 74% achieved by Hierarchical clustering. Furthermore, precision, recall, and F1-score metrics further substantiate the superior performance of K-means. These findings imply that K-means may offer valuable insights for heart disease classification and patient stratification. This work not only contributes to continuing efforts to improve heart disease diagnostics, but it also emphasizes the potential. The study emphasizes the critical role of method selection in data-driven healthcare applications and lays the foundation for future research aimed at optimizing clustering approaches to improve patient outcomes.

Index Terms – **clustering techniques, heart disease, hierarchical algorithm, k-means algorithm**

I. INTRODUCTION

HEART disease remains a prominent global health challenge, accounting for a substantial portion of morbidity and mortality worldwide. As the leading cause of death, its multifaceted nature demands innovative approaches to diagnosis, risk assessment, and treatment. Data-driven methodologies, with their potential to revolutionize our understanding of heart disease, are increasingly important in informing personalized patient care.

The importance of heart disease analysis cannot be overstated. Cardiovascular diseases, which encompass a spectrum of conditions affecting the heart and blood vessels, impose a significant burden on healthcare systems and individuals alike. Accurate and timely diagnosis, informed by treatment decisions, risk stratification, and lifestyle interventions, is pivotal. An in-depth understanding of the heterogeneity within the heart disease population empowers clinicians and researchers to tailor interventions to specific patient profiles.

Clustering techniques, a subset of unsupervised machine learning, have become effective tools for the study of medical data. These approaches have the potential to improve our understanding of disease subtypes, patient classification, and therapy response through the discovery of hidden structures and patterns within large and complicated datasets. Their application in heart disease analysis presents a unique opportunity to identify patient cohorts with distinct clinical characteristics, enabling more precise interventions and better outcomes.

This study focuses on two well-known clustering methods: Hierarchical clustering and K-means clustering. Their utility has been demonstrated in diverse domains, and they have become indispensable in exploratory data analysis. In the context of heart disease classification, they are utilized as powerful tools for the following reasons:

A. Hierarchical Clustering

An agglomerative approach is employed by Hierarchical clustering to build a hierarchical representation of data. It initiates with each data point being treated as its cluster and clusters are subsequently merged based on similarity through recursive processes, resulting in the formation of a tree-like structure known as a dendrogram. This method provides insights into the hierarchical organization of patient subgroups, enabling the identification of nested patterns within the heart disease population.

B. K-means clustering

K-means clustering is a centroid-based technique in which data is partitioned into K clusters, where K is a user-defined parameter. Data points are iteratively assigned to the nearest cluster centroid and centroids are updated until convergence is achieved. Its simplicity and efficiency are known, and it is widely used for grouping patients with similar clinical profiles. The aim is to find compact and well separated clusters within the data.

The primary objective of this study is to assess the efficacy of Hierarchical clustering and K-means clustering in

Y.K.K.M.K. Gayathri is with Dept. of Mathematical Sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanaka (E-mail: minushigayathri@gmail.com).

N.A.D.N. Napagoda, is with Dept. of Mathematical Sciences, Faculty of Applied Sciences, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka (E-mail: namaleen@wyb.ac.lk).

ISSN: 2961-5992

classifying heart disease patients. A rich dataset encompassing a range of clinical attributes is leveraged, enabling a comprehensive analysis of clustering outcomes. The study will delve into precision, recall, and F1-score metrics to provide a nuanced evaluation of each technique's performance. Visual aids, such as dendrograms and cluster plots, will be utilized to enhance the interpretation of results.

This paper seeks to contribute to the burgeoning field of clustering techniques in medical data analysis. The capabilities and limitations of these methods, particularly in the context of heart disease classification, are elucidated to provide insights that may inform clinical decision-making, risk assessment, and the development of tailored treatment strategies. A comprehensive assessment of the clustering strategies used, along with their potential implications for the field of cardiovascular health, is presented through the methodology, results, and discussion in the pages that follow.

II. Literature Review

This literature review explores the application of clustering techniques in medical data analysis, particularly in heart disease classification. The aim is to evaluate the strengths and limitations of various clustering methods, highlight emerging trends, and lay a foundation for the current study's approach to heart disease analysis.

Clustering techniques play a pivotal role in medical data analysis, enabling the discovery of hidden patterns and subgroups within patient populations. These methods have become increasingly significant in heart disease classification due to their ability to reveal valuable insights. A comprehensive review of the literature on Hierarchical clustering and K-means clustering illustrates their applications, comparative studies, and emerging trends in medical data analysis.

Recent advancements have significantly improved the application of clustering techniques in medical data analysis. For example, Hananto et al. [1] demonstrated the effectiveness of the K-means algorithm in drug data analysis, achieving an accuracy rate of 99.45% for condition classification. This study underscores the potential of clustering methods to identify crucial patterns related to drug usage and associated conditions. However, recent breaches involving unauthorized access to drug-related data, with 53,766 cases reported, highlight the urgent need for robust data protection measures to prevent severe consequences, including potential fatalities. Hananto et al.'s [1] work reflects a commitment to enhancing data security and analysis within the healthcare sector.

In terms of methodologies and data preprocessing, significant contributions have been made. Yadav, Tomar, and Agarwal [2] developed the Foggy K-means clustering method for real-time lung cancer datasets, providing more accurate groupings than the traditional K-means approach. Their methodology involved rigorous data preprocessing steps, such as imputation for missing values and feature selection to enhance clustering performance. Similarly, Anas, Gupta, and Ahmad [3] proposed an automated medical image classification method using K-means clustering, utilizing color and texture features to distinguish between melanoma and non-melanoma skin cancer types, significantly improving classification accuracy.

Comparative studies have been crucial in refining our understanding of clustering methods. Ogbuabor and Ugwoke [4] conducted a comparative analysis of K-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) on healthcare datasets, using Silhouette scores. Their study involved varying cluster sizes and distance measures for K-means and different minimum points and distance metrics for DBSCAN. The findings suggest that both methods exhibit strong intra-cluster cohesiveness and inter-cluster separation, with K-means demonstrating superior execution speed and clustering accuracy. Visual aids, such as dendrograms in Hierarchical clustering and cluster plots in K-means, were used to effectively illustrate clustering outcomes.

Emerging trends in data mining reflect the growing volume of data stored in vast databases. Chauhan, Kaur, and Alam [5] explored efficient cluster generation using clustering algorithms on spatial and medical data, revealing previously hidden insights. Singh and Singh [6] conducted a survey focusing on quantized techniques within the medical domain, particularly targeting diabetes susceptibility across different age groups. Despite these advancements, limitations such as data quality and quantity constraints persist, potentially affecting the generalizability of findings. The choice of clustering parameters, such as the number of clusters in K-means, often relies on heuristic methods, which may not always yield optimal results. This study addresses these limitations by using a comprehensive dataset and exploring advanced techniques for parameter selection.

Recent research has further advanced the integration of machine learning with clustering techniques. For instance, Jawalkar et al. [7] proposed a model utilizing K-Modes clustering combined with machine learning classifiers like Random Forest, Decision Tree, and Multilayer Perceptron. This model achieved accuracy rates of up to 87.28% with cross-validation, demonstrating the effectiveness of combining clustering methods with machine learning to enhance diagnostic accuracy. Similarly, Saputra et al. [8] focused on the use of K-means clustering for improving cardiovascular disease prognosis, emphasizing the importance of feature selection and big data integration in personalized medicine.

In the realm of cancer research, recent studies have highlighted the use of Hierarchical clustering for analyzing high-dimensional medical data. Ambigavathi and Sridharan [9] compared various clustering algorithms, including Hierarchical clustering, to identify distinct patient subgroups based on complex biological markers, leading to more personalized treatment approaches.

A broad review of clustering algorithms applied in healthcare underscores the utility of K-means and Hierarchical clustering for processing medical imaging data, disease diagnosis, and prognosis. Saputra et al. [8] demonstrated how these methods enhance decision-making in medical applications by grouping patients based on clinical and physiological data.

In conclusion, clustering techniques, including Hierarchical clustering and K-means clustering, have proven to be valuable tools for uncovering hidden structures within medical datasets. Their application in heart disease analysis has the potential to improve patient stratification, risk assessment, and treatment personalization. Comparative studies and emerging trends continue to advance our understanding of these methods, paving the way for further research in cardiovascular health. This review provides a solid foundation for the current study's exploration of clustering techniques in heart disease classification.

III. Methodology and Experimental Design

A. Data Collection

This paper is based on the Heart Disease Dataset obtained from Kaggle [10]. The dataset, originating from 1988, includes four separate databases: Cleveland, Hungary, Switzerland, and Long Beach V. Although the full dataset contains 76 attributes, this study focuses on a subset of 14 clinically relevant attributes. The attributes used in this study include age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, old peak, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia. The "target" field in the dataset denotes the presence (1) or absence (0) of heart disease.

B. Data Preprocessing

Before clustering analysis, a series of preprocessing steps are applied to ensure data quality and consistency. These preprocessing steps include:

Data Cleaning:

Duplicate records are removed, and missing values are handled using [mean imputation] to ensure data completeness and reliability.

Data Standardization

Features are standardized to have a mean of 0 and a standard deviation of 1. This normalization process ensures that each feature contributes equally to the clustering algorithms and prevents bias due to varying scales.

C. Clustering Techniques

Two clustering techniques, Hierarchical clustering and K-means clustering, are employed to categorize heart disease patients based on their clinical profiles. Each method is applied to the preprocessed dataset as follows:

Hierarchical clustering

It applies Agglomerative Hierarchical clustering with two clusters for binary classification.

K-means clustering

K-means clustering is performed ten times, and the run with the highest Adjusted Rand Index (ARI) is selected as the final K-means result.

D. Quality Evaluation

The Silhouette Score is calculated for both Hierarchical and K-means clustering results to assess their quality.

Cluster Labels Attachment: The cluster labels are attached to the original dataset.

E. Cluster Label Alignment

To ensure K-means cluster labels align with the target variable (0 and 1), they are reversed if necessary.

F. Confusion Matrices

Confusion matrices are calculated for both Hierarchical and K-means clustering results to assess the classification performance. These matrices provide insights into the true positives, false positives, true negatives, and false negatives, facilitating a comprehensive evaluation of each method's effectiveness.

G. Classification Metrics

Various classification metrics, including accuracy, precision, recall, and F1-score, are computed for both clustering techniques. These metrics offer a detailed comparison of the performance of Hierarchical clustering and K-means clustering.

H. Visualization

These visual representations help in understanding the distribution of data points within the clusters and the overall effectiveness of the clustering algorithms.

IV. Results

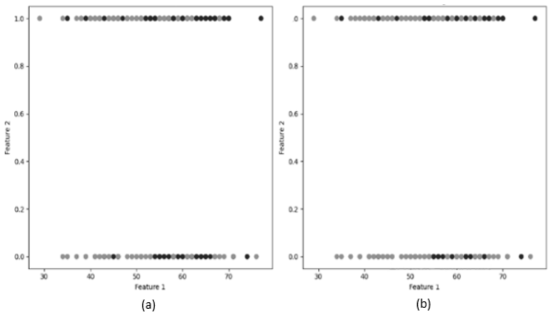

Fig. 1. Visualization of Hierarchical clustering (a) and K-means clustering (b) results for two features.

The scatter plots serve as visual representations of the clustering results, providing insights into how the two clustering methods partition the data. Fig. 1 showcase data points in a two-dimensional space, where each point is color-coded to indicate its cluster assignment. The upper and lower regions in each plot exhibit the data point distribution following hierarchical and K-means clustering. By visually inspecting these plots, one can assess the effectiveness of the clustering algorithms and identify any variations in how they group the data points. This graphical evaluation aids in comparing the clustering outcomes, revealing the extent to which each method separates the data based on the first and second features. This visual examination is particularly useful for understanding the clustering quality and its implications on the dataset's structure.

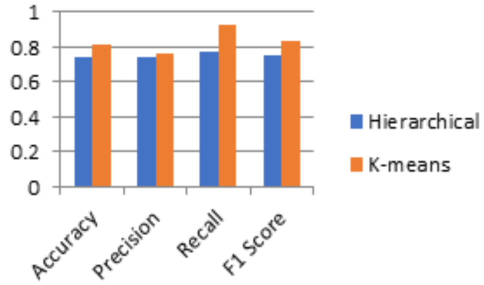

Fig. 2. Bar Chart of Classification Metrics

ISSN: 2961-5992

TABLE I
CLASSIFICATION METRICS RESULTS

	Clustering Algorithms	
	Hierarchical	K-means
Accuracy	0.74	0.81
Precision	0.74	0.76
Recall	0.77	0.93
F1 Score	0.75	0.83

Following the classification metrics presented in TABLE I, Fig. 2 provides a bar chart that visualizes these metrics for both Hierarchical and K-means clustering methods. The bar chart includes metrics such as accuracy, precision, recall, and F1 Score, making it easier to compare the performance of the clustering methods visually. This chart highlights the differences in performance between Hierarchical and K-means clustering across the various evaluation criteria.

The Hierarchical clustering method achieved an accuracy rate of 0.74, while the K-means clustering method outperformed it with an accuracy rate of 0.81, as shown in TABLE I. This indicates that the K-means clustering method achieved a 7% higher accuracy compared to Hierarchical clustering.

In terms of precision, Hierarchical clustering attained a precision rate of 0.74, whereas K-means clustering exhibited a slightly higher precision rate of 0.76. This suggests that K-means clustering had a marginally better ability to correctly classify instances in the positive group.

Regarding recall, Hierarchical clustering achieved a recall rate of 0.77, while K-means clustering demonstrated a notably higher recall rate of 0.93. This highlights K-means clustering's superior ability to capture a larger proportion of positive instances. In the context of the F1 Score, Hierarchical clustering achieved a score of 0.75, while K-means clustering surpassed it with a higher F1 Score of 0.83, as depicted in Fig. 2. This underscores the overall better balance between precision and recall achieved by K-means clustering.

In summary, the results indicate that the K-means clustering method outperformed Hierarchical clustering across multiple evaluation metrics, including accuracy, precision, recall, and F1 Score. These findings demonstrate the superior performance of K-means clustering in classifying the heart disease data.

V. DISCUSSION AND CONCLUSION

In our study, we conducted a comprehensive analysis of heart disease data using two clustering methods: Hierarchical clustering and K-means clustering. The evaluation metrics employed, including accuracy, precision, recall, and the F1 Score, provided insights into the performance of these methods. Our findings indicate that K-means clustering consistently outperforms Hierarchical clustering in all critical metrics. Specifically, K-means demonstrated superior accuracy, precision, recall, and F1 Score, underscoring its effectiveness in grouping heart disease data.

The superior performance of K-means clustering suggests its robustness in uncovering patterns within the data, which could enhance the accuracy of disease classification and support more informed diagnosis and treatment strategies. By achieving higher performance across all evaluation criteria, K-means clustering proves to be a valuable tool for medical data analysis.

The results highlight the potential of K-means clustering to improve insights into heart disease patterns, which may have significant implications for medical practice. By leveraging K-means, healthcare professionals could gain a more nuanced understanding of patient subgroups, enabling more personalized and effective treatment approaches.

While this study emphasizes the advantages of K-means clustering, several avenues for future research are suggested:

Incorporation of Additional Features: Future studies could explore the inclusion of additional clinical features or advanced preprocessing techniques to further refine clustering performance and enhance the quality of insights.

Comparison with Other Clustering Algorithms: Investigating other clustering algorithms, such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) or spectral clustering, may provide additional perspectives on clustering efficacy in the context of heart disease data.

Development of Predictive Models: Building predictive models based on clustered data could aid in assessing disease risk and improving early diagnosis.

Clinical Validation:Collaborating with medical experts to validate the clinical relevance of identified clusters and their potential impact on patient care will be crucial. Such collaboration could ensure that clustering results translate into actionable clinical insights and benefits for patient management.

In conclusion, our study confirms the efficacy of K-means clustering in the analysis of heart disease data and opens several pathways for future research to further explore and validate the potential of clustering techniques in medical data analysis.

REFERENCES

[1] A. L. Hananto, P. Assiroj, B. Priyatna, A. Fauzi, A. Y. Rahman, and S. S. Hilabi, "Analysis of drug data mining with clustering technique using K-Means algorithm," J. Phys. Conf. Ser., vol. 1908, no. 1, 012024, 2021. Available at: https://doi.org/10.1088/1742-6596/1908/1/012024. [Accessed: Sep. 9, 2024].

[2] A. K. Yadav, D. Tomar, and S. Agarwal, "Clustering of lung cancer data using foggy K-means," in 2013 International Conference on Recent Trends in Information Technology (ICRTIT), 2013, pp. 13-18.

[3] M. Anas, K. Gupta, and S. Ahmad, "Skin cancer classification using K-means clustering," Int. J. Technical Res. Appl., vol. 5, no. 1, pp. 62-65, 2017.

[4] G. Ogbuabor and F. N. Ugwoke, "Clustering algorithm for a healthcare dataset using silhouette score value," Int. J. Comput. Sci. Inf. Technol., vol. 10, no. 2, pp. 27-37, 2018.

[5] R. Chauhan, H. Kaur, and M. A. Alam, "Data clustering method for discovering clusters in spatial cancer databases," Int. J. Comput. Appl., vol. 10, no. 6, pp. 9-14, 2010.

[6] G. Singh and G. Singh, "Diabetes classification using K-Means," APEJAY J. Comput. Sci. Appl., 2019.

[7] A. P. Jawalkar et al., "Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting," J. Eng. Appl. Sci., vol. 70, no. 1, p. 122, 2023.

[8] J. Saputra, C. Lawrencya, J. M. Saini, and S. Suharjito, "Hyperparameter optimization for cardiovascular disease data-driven prognostic system," Visual Comput. Ind., Biomed., and Art, vol. 6, no. 1, p. 16, 2023.

[9] M. Ambigavathi and D. Sridharan, "Analysis of clustering algorithms in machine learning for healthcare data," in Advances in Computing and Data Sciences: 4th International Conference, ICACDS 2020, Valletta, Malta, April 24–25, 2020, Revised Selected Papers 4, M.

ISSN: 2961-5992

Singh, P. Gupta, V. Tyagi, J. Flusser, T. Ören, and G. Valentino, Eds. Springer Singapore, 2020, pp. 117-128.

[10] Kaggle, "Heart Disease Dataset," available at: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset. [Accessed: Sep. 9, 2024].

ISSN: 2961-5992