International Journal of Innovation and Multidisciplinary Research
DOI.ORG/10.67505/IJIAMR.v6.i2.62001

Vol. 6 (2), 2026
E-ISSN: 2583-4452

A Comparative Analysis of Clustering Algorithms in Targeted Preventive Care

Shalvi Tripathi[1*], Garima Singh[2], Arushi Goyal[3]

[1, 2, 3] Department of Mathematics, Shaheed Rajguru College of Applied Sciences for Women, University of Delhi, Delhi

[*] Corresponding Author ✉ jagatpoonam7303@gmail.com

Abstract

In the modern healthcare, the importance of preventive care has emerged as an effective approach in order to lighten the load of hospitals while also ensuring prolonged life span for patients. The current study on Clustering algorithms in targeted preventive care, highlights one such major issue in clustering algorithms that help uncover hidden patterns from a dataset to generate more precise preventative actions in medical care. The core of this study involves a detailed comparative study of different clustering algorithms: K-Means, DBSCAN and Hierarchical Clustering. The methodology used in this study involves using various health datasets containing such features as age, body mass index, diabetes, asthma, arthritis, life style factors and geographical location, etc. The main evaluation factors to consider in regard to the algorithms include Silhouette scores metric, Scalability, Robustness and Interpretability. Based on the results of this analysis, K-Means proves to be very efficient when dealing with huge, well-defined datasets. For this reason, it is suitable for population screening. However, when considering identification of outliers in patients having rare or unusual health problems, DBSCAN shows much better results. On the other hand, Hierarchical clustering gives much better insight into nesting various health risks. This report also focuses on preprocessing, including scaling of data and handling incomplete values. Another interesting point of the analysis is that besides simple classification of health data, it may help predict chronic diseases, discovering some symptoms that would remain unnoticed during manual examination.

Keywords: *K-means, DBSCAN, Hierarchical Method, Healthcare Analytics, Patient Segmentation*

1. Introduction

Today, healthcare systems generate vast volumes of patient data from electronic medical records, laboratory reports, clinical observations, and lifestyle information. While this data holds significant potential for improving healthcare outcomes, its high dimensionality, complexity, and heterogeneity make effective analysis challenging. Without appropriate analytical techniques, it becomes difficult to identify hidden patterns, stratify patients based on risk levels, and support early preventive interventions. Therefore the need of clustering algorithm arises due to its potential to find early disease detection, lower healthcare costs, optimize population well-being and timely intervention. Clustering algorithms are techniques that automatically group data into classes based on similarities but without predefined labels. Each of the clustering algorithms follows a unique way of working based on different assumptions. Therefore, there is variation in their working and performance. Some of the common clustering algorithms used in the field of health care include K-Means, DBSCAN, and hierarchical clustering. These algorithms are widely used in classifying patients in accordance with the similarities found in the health care data. However, there are no such set criterion to determine what algorithm would be best for preventative healthcare. Most of the healthcare related clustering studies conducted so far have employed only one kind of clustering algorithm. To address this gap, the present study evaluates and compares three different methods to provide a more detailed and balanced analysis. The primary goal of the current study is related to conducting a comparative analysis of the three clustering algorithms which include K-Means, DBSCAN, and Hierarchical Clustering used in classifying patients within preventive healthcare. The methodology includes selection of a healthcare dataset, data preprocessing, implementation of different clustering algorithms, and evaluation of their performance using appropriate metrics such as scalability, robustness, interpretability, and medical relevance. This differs from other studies whose emphasis lies on accuracy. However, despite the existence of many articles comparing clustering algorithms, few works have been concerned about their performances such as the accuracy, silhouette coefficient, or robustness while the other aspect like the interpretability remains ignored. Therefore, an opportunity is opened where interpretability is essential to get insight into the patient's risk pattern and make clinical decision accurately. In order to fill this gap, this study conducts the comparison of different clustering algorithms on a large healthcare dataset of 20,000 patient records by four criteria, among which there is the interpretability one. This paper makes a unique contribution to the field of medical data clustering by taking into account both aspects of performance and interpretability. The necessity of this paper comes from the need of patient stratification in advance to conduct healthcare intervention. The

Received on: 10 June 2026 | Accepted on: 27 July 2026 | Published Online on: 20 November 26

© 2026 The Author(s). This article is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0).

performance of these clustering algorithms in dealing with data which are large, noisy, and multidimensional, as well as the strengths and weaknesses, is examined in order to bridge the gap between theory and application. In addition, another equally important goal of this study is to develop an organized framework for comparing different types of clustering approaches. Instead of randomly selecting algorithms by using trial and error approaches, this study highlights the conditions where a certain algorithm should be applied. Many researches on clustering methods have been done in great depth and have been used in various fields including health care and data analysis, among others. At the initial stages, many researchers have tried to develop the fundamental clustering approaches. One of the important contributions was given by Celebi et al. [7] who examined different initialization methods for K-means and showed that the initial stage of clusters can significantly affect the final outcomes. Sinaga et al. [20] introduced an improved version of K-means, known as unsupervised K-means, which can independently identify the required number of groups without requiring user input. Clustering algorithms have been broadly employed within the healthcare domain to improve patient analysis and decision-making. Taiwo et al. [22] proposed using the K-means algorithm to cluster high-risk patients based on their health status, helping to develop efficient preventive measures. In addition, Ghazal et al. [11] studied the impact of different distance measures on clustering efficiency and found that the selection of distance measure plays an important contribution in both accuracy and execution time. Apart from the cluster-based approaches that are based on centroid, density-based algorithms such as DBSCAN have also been thoroughly explored. DBSCAN was introduced by Ester et al. [9]. As time went on, a number of researchers began looking into ways to enhance the existing algorithm's weaknesses. For instance, Aljibawi et al. [3] introduced the idea of an adaptive DBSCAN (ADBSCAN) algorithm capable of self-regulating according to the data in question, making it more effective when dealing with non-uniformly distributed data containing noise. Another important aspect of clustering research is assessment of clustering outcomes. Ogbuabor and Ugwoke [17] highlighted the importance of using Silhouette Score to measure the quality of groups and to select the optimal number of clusters. Their work shows that choosing an incorrect number of groups can produce to poor outcomes, while the correct choice improves both interpretability and accuracy. Additionally, according to Andreopoulos [5], hierarchical and k-means are some of the widely used clustering methods. The work evaluates 40 different clustering methods, stating the importance of matching methods to biomedical application needs. This research provides a comparison of different algorithms, based on their assessment measures and the advantages and disadvantages they carry. There are also other studies that concentrate on certain trends and concepts associated with data analysis. For instance, in their research Oyewole and Thopil [18], Neeraj and Maurya [16] covered different clustering approaches as well as feature selection algorithms. In their research, authors pointed out some trends and development associated with big data analytics as well as machine learning-based data mining. Thus, they demonstrated how all these concepts together can help in enhancing the performance of data analysis. Another interesting study by Hasibuan et al. [13] concentrated on a comparative study of different optimization algorithms. Specifically, authors analyzed the efficiency and convergence rate of algorithms.

This research offers important insights into the choice of optimization approaches for different types of computations. Numerous researchers have also presented review papers to present and compare diverse clustering algorithms. In particular, Ezugwu et al. [10] made an extensive review of both traditional and contemporary clustering techniques, paying significant attention to their core concepts and problems related to their use with complicated data sets. In addition, Wani et al. [23] reviewed different clustering techniques and moved to discuss more sophisticated approaches, such as deep clustering and ensembles, which become more popular with respect to large amounts and higher dimensions of data. Ahmad and Khan [2] concentrated exclusively on clustering algorithms applicable to mixed data containing numeric attributes and categorical ones.

These studies prove that clustering techniques are very useful in patient segmentation, predicting disease prevalence, identifying risks, and helping to make decisions in the healthcare field. However, most researchers consider particular clustering techniques and find out that no clustering technique works well in every case. As it is noted by the authors under review, the effectiveness of clustering techniques is affected by many factors including the type of healthcare data, data preprocessing, accuracy, speed, stability, scalability, robustness, and interpretability of the results. While some techniques work effectively with structured and large volumes of data, others prove to be helpful in dealing with noisy and complex data sets. That is why these studies emphasize the need for comparative analysis of different clustering techniques.

2. Different Clustering Approaches and Evaluation Parameters

2.1 Clustering Algorithms

Clustering algorithms help find hidden patterns in big datasets, which is really useful in healthcare. They group similar data points together to make analysis easier. This study looks at several clustering methods to find patient groups, aiming to boost the efficiency and accuracy of decisions made from data. We check out methods like K-Means, DBSCAN, and Hierarchical Clustering. To evaluate them, we use criteria like the silhouette score, how easy they are to understand, how robust they are, and whether they can handle lots of data without issues.

2.1.1 K-Means Clustering

K-Means Clustering refers to an important unsupervised learning method that is extensively employed in data science and machine learning for dividing data points into clusters based on similarity. The objective of K-Means Clustering is to divide the given n number of data points into k clusters where each data point should be placed in the cluster that has the closest centroid. It works by minimizing the intra-cluster variance and thereby forming groups with similar data points. [1], [4]

The goal of K-Means is to find clusters where sum of squared error is minimized

$$J = \sum_{i=1}^{k} \sum_{x \in c_i} \left\| x - \mu_i \right\|^2 \tag{1}$$

Where:

- C_i represents collection of points belonging to cluster i
- μ_i represents centroid for cluster i
- $||x - \mu_i||^2$ denotes square of Euclidean distance between data point x and centroid μi

The approach aims at reducing the total variability within each cluster, which is known as inertia and defined as the sum of the squared differences between each observation and its cluster center. K-means results in a set of K cluster centers with each observation assigned to the nearest cluster center.

2.1.2 DBSCAN Clustering

DBSCAN algorithm is developed by Martin Ester, Jorg Sander, Xiaowei Xu, and Hans-Peter Kriegel. The DBSCAN algorithm uses density-based clustering method and forms clusters based on how dense are the data points in the feature space. The partition-based clustering algorithms form clusters based on density. However, DBSCAN forms clusters as contiguous regions with high density points separated from other regions by regions of lower densities. If an epsilon-neighborhood of a point p contains at least MinPts points then that point is a core point. [1], [4]

$$d(p,q) = \sqrt{\sum_{i=1}^{n} \left(q_i - p_i\right)^2} \tag{2}$$

The ϵ-neighborhood of a point p consists of points no farther than ϵ away from p.

$$N_\epsilon(p) = \{q \in D \mid d(p, q) \leq \epsilon\} \tag{3}$$

A point p is a core point if its ϵ-neighborhood contains at least MinPts points.

$$|N_\epsilon(p)| \geq \text{MinPts} \tag{4}$$

Where:

- $|\cdot|$ represents the cardinality (size of) a set.
- MinPts is the minimum number of points needed to constitute a dense region.

DBSCAN works very well in detecting cluster formations with complex forms and dealing with noises in the data set. This makes DBSCAN a reliable technique for practical applications. Different from many conventional clustering methods, DBSCAN does not need the exact number of clusters to be predefined, thus increasing its adaptability to practical uses. Nevertheless, the application effect of DBSCAN is dependent on the correct setting of the core parameters, ϵ (epsilon) and MinPts. In spite of this difficulty, the well-set DBSCAN can be regarded as an effective approach for complex patterns detection in spatial data sets.

2.1.3 Hierarchical Clustering

Hierarchical clustering is the third learning paradigm in the domain of unsupervised learning, which is characterized by its connectivity-based strategy for creating a nested series of clusters. These techniques are typically grouped into two broad categories, the Agglomerative clustering method, which initiates with each datum as a separate cluster and continues merging the two closest clusters until there is only one cluster left, and the Divisive clustering algorithm, where all data points are initially assigned to one cluster, and the clusters are subsequently subdivided until each cluster contains a single point. [1],[4]

$$D(C_i, C_j) = \text{linkage function } (\{d(p, q) \mid p \in C_i, q \in C_j\}) \tag{5}$$

Linkage Function: Ward's Method:

$$D_{Ward}\left(C_i, C_j\right) = \frac{|C_i||C_J|}{|C_i| + |C_J|} \left\| \mu_i - \mu_j \right\|^2 \tag{6}$$

In which:

- $|C_i|$ and $|C_j|$ represent the sizes of the clusters C_i and C_j.
- μ_i and μ_j represent the centroids (means) of the clusters C_i and C_j.

Hierarchical clustering is most convincing strength in terms of patient stratification when it comes to preventative medicine is undoubtedly the ease with which the results lend themselves to clinical interpretation as well as hierarchical complexity. The output that arises from Hierarchical Clustering analysis is the dendrogram, which is a classic tree diagram that traces the entire history of mergers, with distances (or dissimilarities) represented on the y-axis at which such mergers took place. This allows for a deep and meaningful insight into the hierarchy of patient sub-groups.

2.2 Evaluation parameters

The effectiveness of clustering models is judged according to their cluster quality and interpretability so that the clusters are not only useful but understandable from a medical point of view. Moreover, simulations are performed for the comparison between targeted and non-targeted preventive care strategies to show how clustering models can facilitate the development of better healthcare interventions.

2.2.1 Cluster Quality Metrics

Cluster quality measures are metrics employed to determine how well a particular clustering algorithm can group data that belongs in a cluster, while at the same time ensuring that the data in the cluster is clearly distinguishable from any other cluster. The silhouette coefficient is an extensively applied method to determine the effectiveness of clustering outcomes based on the degree of coherence within a cluster as opposed to other clusters. [17]

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{7}$$

Where:

- $a(i)$ is the average distance between data item i and other items within the same cluster.
- $b(i)$ is the minimum average distance between data item i and all other items in the nearest neighboring cluster.

This score varies between -1 and +1. Scores closer to +1 mean that the clusters formed are distinctly separate, scores near zero represent overlapping clusters while negative values mean that the wrong clusters have been formed. Because of its capability to balance the two factors, the silhouette coefficient is popularly used for evaluating various clustering methods.

2.2.2 Robustness

Robustness refers to the characteristic of the clustering method's ability to generate consistent and reliable outputs regardless of fluctuations in the input data such as the introduction of noise, anomalies, or even minor perturbations in the dataset itself. The robustness of various algorithms can be tested through their reaction to input data fluctuations. [6]

$$Robustness = \arg\min \| \delta \| \tag{8}$$

Where:

- δ denotes the least amount of noise required to alter the output of the model.

$$\delta = |S - S'| \tag{9}$$

Where,

- S is the silhouette score of the unmodified dataset
- S' is the silhouette score after applying noise

This feature becomes especially significant when implementing such algorithms practically, as the input may not be perfect in many cases. Algorithms that are more robust are resistant to any noise and variations in initial conditions, thus making their performance more reliable.

2.2.3 Scalability

The concept of scalability refers to the capability of clustering algorithms to cope effectively with larger amounts of data without losing efficiency or accuracy. [21]

$$Scalability \propto \frac{1}{Execution\,Time\,Growth} \tag{10}$$

The scalability of an algorithm ensures that it is possible to find relevant clusters and patterns despite the increase in the volume of data. The greater the scalability of an algorithm, the more suitable it will be for implementation in a healthcare system because new data are continuously generated.

2.2.4 Interpretability

Interpretability is about how comprehensible and interpretable the output produced by the clustering algorithms are to human beings, especially when used in a clinical context. The interpretability of a clustering method will be assessed by examining the clarity of interpretation that the clustering algorithm provides regarding the data clusters. [14]

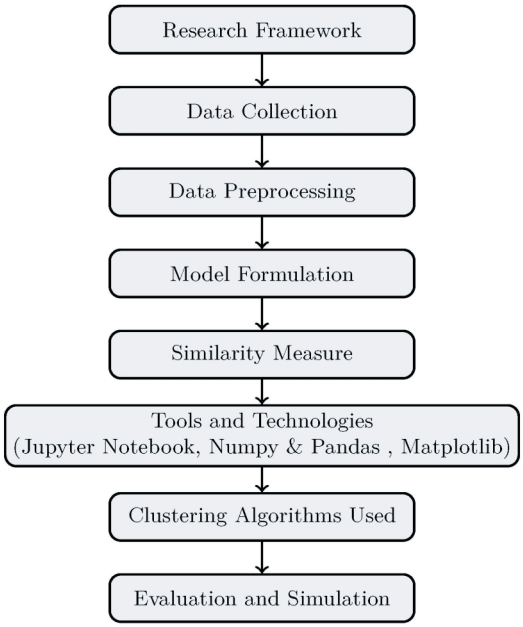
$$Interpretability = \frac{1}{2}\left(C_s + F_s\right) \tag{11}$$

- C_s = Cluster separation measure (silhouette measure)
- F_s = Feature simplicity measure (clusters explainability)

Interpretability helps in hierarchical clustering, the utilization of a dendrogram assumes great significance because it helps visualize the relationships between clusters and the order in which they merge. This facilitates understanding the hierarchical organization of clusters and helps choose the correct number of clusters.

3. Methodology

In order to accomplish the aims of the study, a systematic process was undertaken. The process comprises selection of healthcare data, pre-processing, use of different clustering methods, and assessing their effectiveness through certain criteria. This scientific process plays an important role in evaluating the effectiveness of clustering methods in preventive healthcare analysis.



3.1 Research Framework

This framework uses a Multi-Criteria Decision Making (MCDM) approach to enhance patient stratification [4]. Firstly, a big size healthcare dataset is gathered. Since this raw data cannot be analyzed directly due to its complexity, the preprocessing of the dataset has been conducted. Furthermore, normalization techniques such as Min-Max scaling has been used in this step. After preprocessing, PCA has been applied in order to conduct a necessary dimensionality reduction process and deal with high-dimensional data also discussed by Malik [15]. Afterward, preprocessed and normalized data has been used as input for further clustering procedures. In the next stage, three types of clustering algorithms: K-Means, DBSCAN, and hierarchical clustering (Ward method) has been used to cluster patients. The basis for measuring patient similarity within the framework of this paper has used Euclidean distance, which can provide an opportunity to form clusters of similar patients. Finally, the results are obtained by applying these methods and are further analysed with the use of validation indices such as Silhouette Score. Additional factors like scalability, robustness, and interpretability are also considered to assess the overall effectiveness of each algorithm.

3.2 Data collection

In this study, secondary data has been used. The data is collected from published source available on the web which is owned by Tarek Muhammed Abdel-Hamid and viewed by Kamil Pytlak for their research on Heart Disease Prediction which is related to healthcare [8]. The data consists of patient information with 237630 observations and 35 variables. These variables include age, gender, state, general health, BMI, and many more that relate to health issues and demographics. Diseases such as heart attack, diabetes, stroke, asthma, and other health issues have been captured in the dataset. The information also includes lifestyle aspects like smoking, drinking, and vaccination status as shown in Table 1.

3.3 Data Preprocessing & Cleaning

The dataset that was collected cannot be put to use directly for the purpose of execution and analysis as there are many problems with the quality of data. There are missing values in certain features, which may cause incomplete or inaccurate results if not addressed. Also, there are instances where in non-numeric entries exist in some features when numeric entries should have been present. These undesirable characteristics contributes to increased computational complexities, cause noise, adversely impacting

the effectiveness of the clustering algorithms. Thus, it is crucial that proper preprocessing steps be undertaken prior to analyzing the dataset. The preprocessing methods used in this study were motivated by the works of Aggarwal [1] & Rodriguez [19], where they highlight the significance of preprocessing and propose a more systematic way of evaluating clustering algorithms. The raw dataset was preprocessed to enhance its quality and ensure consistency for analysis. Data cleaning techniques like dealing with missing values, eliminating noise, and normalization of features were applied. As a result, a refined and structured dataset was obtained, which is shown in Table 2.

Table 1: *Dataset with Demographic, Clinical, and Lifestyle Attributes of Patients*

ID	State	Sex	Age Cat.	Height (m)	Weight (kg)	BMI	Angina	Asthma	Skin CA	C OPD	Kidney	Arthritis	Diabetes	Smoker	E-cig	CT	Alcohol	HIV	Flu	Pneumo	Covid
1	Alabama	Female	75–79	1.63	84.82	32.10	1	1	1	0	0	1	Yes	Former	Never	1	0	0	0	1	1
2	Alabama	Female	65–69	1.60	71.67	27.99	0	0	0	0	0	1	No	Former	Never	0	0	0	1	1	0
3	Alabama	Male	60–64	1.78	71.21	22.53	0	0	0	0	0	0	No	Never	Never	0	1	0	0	0	0
4	Alabama	Male	70–74	1.78	95.25	30.13	0	0	0	0	0	1	Yes	Former	Never	0	0	0	1	1	0
5	Alabama	Female	50–54	1.68	78.02	27.76	0	0	0	0	0	0	No	Never	Never	1	0	0	1	0	0
6	Alabama	Male	75–79	1.85	108.86	31.66	0	0	0	0	0	1	No	Former	Never	1	1	0	0	1	1
7	Alabama	Female	70–74	1.70	74.84	25.84	0	0	0	0	0	1	No	Never	Never	1	1	0	0	0	0
8	Alaska	Female	60–64	1.70	87.54	30.23	0	0	0	0	0	1	Pre-DM	Never	Never	0	0	0	0	1	0
9	Alabama	Female	80+	1.70	90.72	31.32	0	0	1	0	0	1	No	Never	Never	0	0	0	1	1	1
10	Alabama	Female	80+	1.55	79.38	33.07	0	0	0	0	0	1	No	Never	Never	0	0	0	1	1	0
11	Alaska	Female	55–59	1.60	113.40	44.29	0	0	0	1	0	0	No	Never	Never	0	0	0	0	0	0
12	Alaska	Male	25–29	1.80	79.83	24.55	0	0	0	0	0	0	No	Never	Current	0	0	0	0	0	0
13	Alabama	Male	50–54	1.85	120.20	34.96	0	0	0	0	0	0	No	Never	Never	1	1	0	1	0	0
14	Alabama	Female	60–64	1.65	57.61	21.13	0	0	0	0	0	0	No	Current	Current	1	1	0	0	0	1
15	Alabama	Female	80+	1.63	88.00	33.30	0	0	1	0	0	0	Yes	Never	Never	1	0	0	1	1	0
16	Alabama	Male	65–69	1.70	90.72	31.32	0	0	0	0	0	0	No	Never	Never	0	0	0	0	0	0
17	Alabama	Male	75–79	1.75	74.84	24.37	1	1	0	0	1	0	Yes	Never	Never	1	0	0	1	1	1
18	Alabama	Female	75–79	1.57	74.84	30.18	0	1	0	0	1	1	No	Former	Never	0	1	1	0	1	0
19	Alabama	Male	25–29	1.73	72.57	24.33	0	0	0	0	0	0	No	Never	Never	0	0	0	0	0	0
20	Alabama	Male	40–44	1.70	78.02	26.94	0	0	0	0	0	0	No	Former	Never	1	0	0	0	0	1
21	Alabama	Male	80+	1.73	91.17	30.56	0	0	1	0	0	0	No	Never	Never	1	0	0	1	1	0
22	Alabama	Female	75–79	1.68	63.50	22.60	0	0	0	1	1	0	No	Former	Never	1	0	0	1	1	0
23	Alabama	Female	30–34	1.55	83.46	34.77	0	0	0	0	0	0	No	Never	Never	0	0	0	0	0	1
24	Alabama	Male	75–79	1.80	102.06	31.38	0	0	0	0	1	0	No	Never	Never	1	0	0	0	1	1
25	Alabama	Male	70–74	1.73	86.18	28.89	1	1	0	1	0	0	Yes	Never	Never	1	0	0	0	1	1

Table 2: *Dataset after Preprocessing*

PatientID	Height	Weight	BMI	Angina	Stroke	Asthma	SkinCancer	COPD	Depressive	Diabetes(yes)	Diabetes(No),pre -diabetes	Diabetes(yes) but only during
1	1.63	84.82	32.10	1	0	1	1	0	0	True	False	False
2	1.60	71.67	27.99	0	0	0	0	0	0	False	False	False
3	1.78	71.21	22.53	0	0	0	0	0	0	False	False	False
4	1.78	95.25	30.13	0	0	0	0	0	0	True	False	False
5	1.68	78.02	27.76	0	0	0	0	0	0	False	False	False
6	1.85	108.86	31.66	0	0	0	0	0	0	False	False	False
7	1.7	78.02	27.76	0	0	0	0	0	0	False	False	False
9	1.7	78.02	27.76	0	0	0	0	0	0	False	False	False
10	1.55	79.38	33.07	0	0	0	0	0	0	True	False	False
11	1.60	113 .40	44.29	0	0	0	1	0	0	False	False	False
12	1.80	79.83	24.55	0	0	0	0	0	0	False	False	False
13	1.85	120.20	34.96	0	0	0	0	0	0	False	False	False
14	1.65	57.61	21.13	0	0	0	0	0	0	False	False	False
15	1.63	88.00	33.30	0	0	1	0	0	0	False	False	False
16	1.70	90.72	31.32	0	0	0	0	0	0	True	False	False
17	1.75	74.84	24.37	1	1	0	0	1	0	True	False	False
18	1.57	74.84	30.18	0	1	0	0	1	1	True	False	False
19	1.73	72.57	24.33	0	0	0	0	0	0	False	False	False
20	1.70	78.02	26.94	0	0	0	0	0	0	True	True	False
21	1.73	91.17	30.56	0	0	1	0	0	0	False	False	False
22	1.68	63.50	22.60	0	0	0	1	1	0	True	False	False
23	1.55	83.46	34.77	0	0	0	0	0	0	False	False	False
24	1.80	102.06	31.38	0	0	0	0	1	0	True	False	False
25	1.73	86.18	28.89	1	1	0	1	0	0	False	False	False

3.4 Model Formulation

The problem of patient segmentation in this study is formulated as an unsupervised learning problem. The goal of the task is to group the patients in clusters based on their similarity, which does not require predefined labels. Assume that the set of our input data can be expressed as follows: $D = \{x_1, x_2, x_3, \ldots, x_n\}$, where each individual entry x_i is a feature vector describing a certain patient. Each patient is mapped to a feature vector in a d-dimensional space, i.e., $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$. The Euclidean distance is used for finding the similarity among any two patients. Patients are grouped together based on the similarity measures such that the similarity among patients within each cluster is maximized and minimized among different clusters.

3.5 Similarity Measure

Harmouch et al. has described many similarity measures and concluded that there are 17 methods available for computing data similarity [12]. Among those 17 similarity measures, Euclidean distance is among the simplest yet effective methods for measuring similarities. Euclidean distance is used in this research work for finding clusters of patient groups. Euclidean distance finds out the direct distance between two data points located in multidimensional space. Distance between two data points can be defined as:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}\left(x_i-y_i\right)^2} \tag{12}$$

Here, x and y represent two data points, and n is the number of features.

3.6 Tools and Technologies

The proposed implementation of this study involves the use of the Jupyter Notebook framework, which offers a live programming environment to code, visualize, and analyze the process. The languages that are employed for coding are Python with Pandas and Numpy libraries for data preprocessing, as well as Matplotlib for visualization. Data preprocessing methods involve data cleaning, managing missing data points, choosing features, and conducting Min-Max normalization. In addition, the principal component analysis method is used for dimensionality reduction.

3.7 Clustering Algorithms

The data gets analyzed using three different techniques. First, K-Means Clustering, which splits the dataset into a set number of groups based on similarities. Then, DBSCAN, which finds clusters of all shapes by looking at data density. It also spots any noise or outliers in the data. Lastly, Hierarchical Clustering, which creates a tree of clusters showing how data relates at various similarity levels.

3.8 Evaluation and Simulation

The clustering model's performance depends on their cluster quality and how easy those clusters are to understand, making them meaningful in a clinical way. Also, we run simulations comparing targeted and non-targeted preventive care. This shows how clustering helps create more effective, personalized health interventions.

4. Results and Discussion

Upon running the algorithms, all three gave clustering outputs that were unique in nature because of the methods used by each algorithm. These clustering outputs are important as they give a good understanding of the distribution pattern within the dataset and offer different views regarding groupings of the data points. This makes it easier for one to get a good understanding of what each algorithm can do in terms of capturing the distribution within the dataset. These outputs have been discussed in detail below.

4.1 K-Means

In K-Means clustering analysis, the number of clusters (k) has to be specified before proceeding with any other step. However, choosing an incorrect number for k may result in clustering being either ineffective, owing to the formation of too many clusters, or ineffective, because some clusters would be merged into one. Thus, a technique must be devised to find out the right number of clusters. This problem has been solved using the elbow method, and its graph has shown in Figure 1. In this technique, we have used the K-Means algorithm for several k values and calculate the Within-Cluster Sum of Squares (WCSS) for each one of them. WCSS represents the closeness of data within clusters, which means smaller WCSS results in better clustering. For increasing values of k, WCSS decreases gradually. But, beyond a particular point, the slope of this decrease becomes very gentle. That particular point is called the "elbow point".

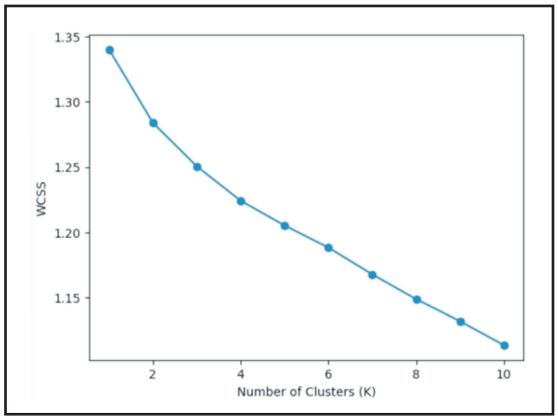


Figure 1 *:Elbow Curve for optimal value of K*

From the Elbow plot analysis, it showed that the elbow is obtained when k=5, implying that there will be no gain obtained from increasing the cluster number above this level since clustering performance will not improve. Therefore, the optimal number of clusters for the K-means Algorithm is k=5. The application of the K-Means Algorithm on the dataset with the optimal number of

clusters (k=5) yields an efficient classification into five classes. The figure below shows the 3D visual presentation of the clustered data as shown in Figure 2.

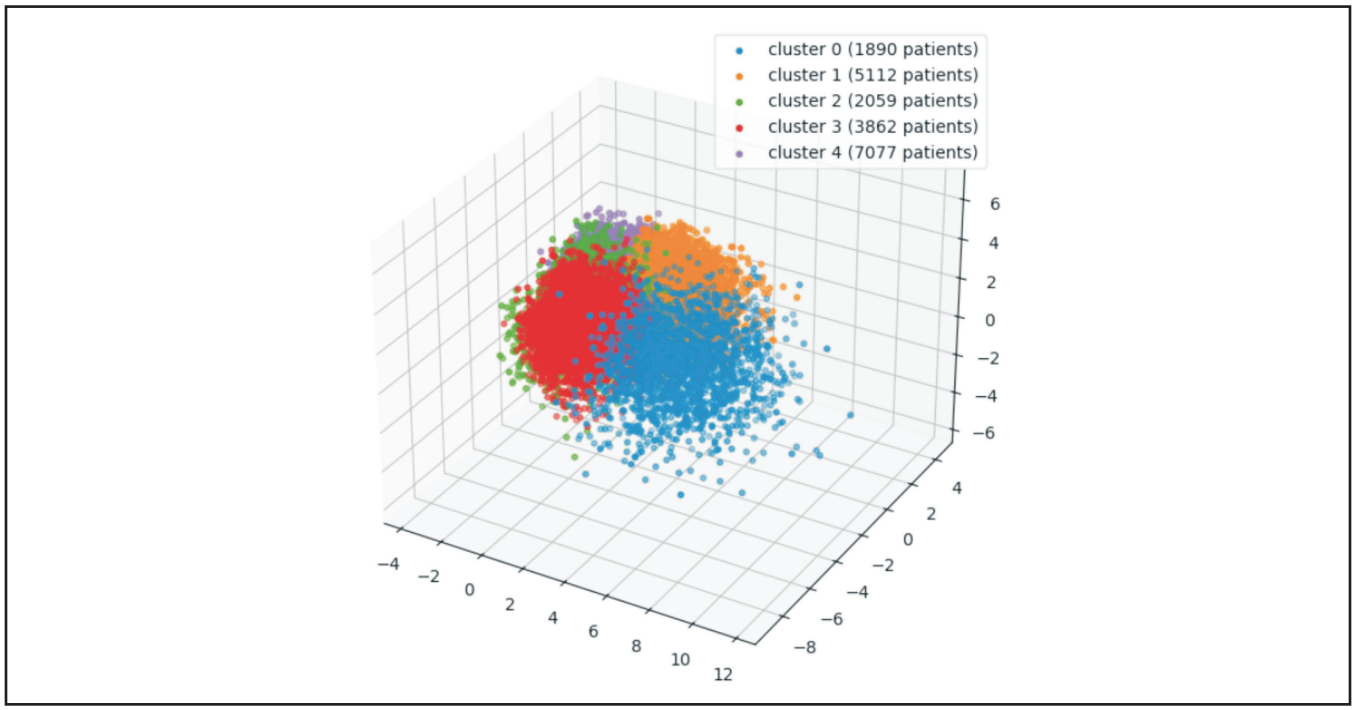


Figure 2: *K-Means Clustering of Patient Data*

The algorithm successfully clustered the data into appropriate clusters depending on similarities within the data. Each cluster is a collection of people who share similar traits in the projected feature space. There is clear separation and grouping of these clusters showing that the algorithm has successfully identified patterns within the data set. The variation in cluster sizes suggests heterogeneity in the dataset, with some patient groups being more prevalent than others. Overall, the clustering result demonstrates meaningful segmentation of the dataset, which can further support subgroup analysis and interpretation of patient profiles.

4.2 DBSCAN Clustering

The DBSCAN Clustering was perfomed on the dataset to identify clusters according to the density in place of pre-defined cluster counts. In DBSCAN, an epsilon value of 10 was selected as it provided the best grouping efficiency in concept of the silhouette score. When the epsilon value was reduced, the silhouette score decreased significantly, indicating weaker cluster quality and less meaningful separation among data points. Although an epsilon value of 3 also produced a relatively high silhouette score, it resulted in a large number of points (approximately 40 percent of the dataset) being classified as outliers, which makes it unsuitable for meaningful clustering in Table 3.

Table 3: *Cluster count with different ϵ values*

Epsilon (ϵ)	Unique Clusters	Cluster Count
0.5	{-1}	1
0.7	{-1}	1
1.0	{0, -1}	2
1.5	{0, -1}	2
2.0	{0, -1}	2
2.5	{0, 1, 2, ..., 16, -1}	18
3.0	{0, 1, 2, ..., 45, -1}	47
8.0	{0, 1, -1}	3
10	{0, 1, -1}	3
11	{0, 1, -1}	3
12	{0, -1}	2

Based on this analysis, the appropriate number of clusters was determined. Subsequently, Silhouette Score has been computed for the optimal grouping result that evaluate the quality & cohesion of the formed groups. The results obtained for different epsilon values, along with the corresponding number of clusters, are presented in Table 4.

Table 3: *Cluster count with different ϵ values*

ϵ	No. of Cluster	Score
3	47	0.3442
8	3	0.3138
10	3	0.3088
11	3	0.3099

Therefore, ϵ = 10 was chosen as the optimal value as it achieves a better balance between cluster formation and noise handling, along with a higher overall silhouette score, which produced three clusters as shown in Figure 3.

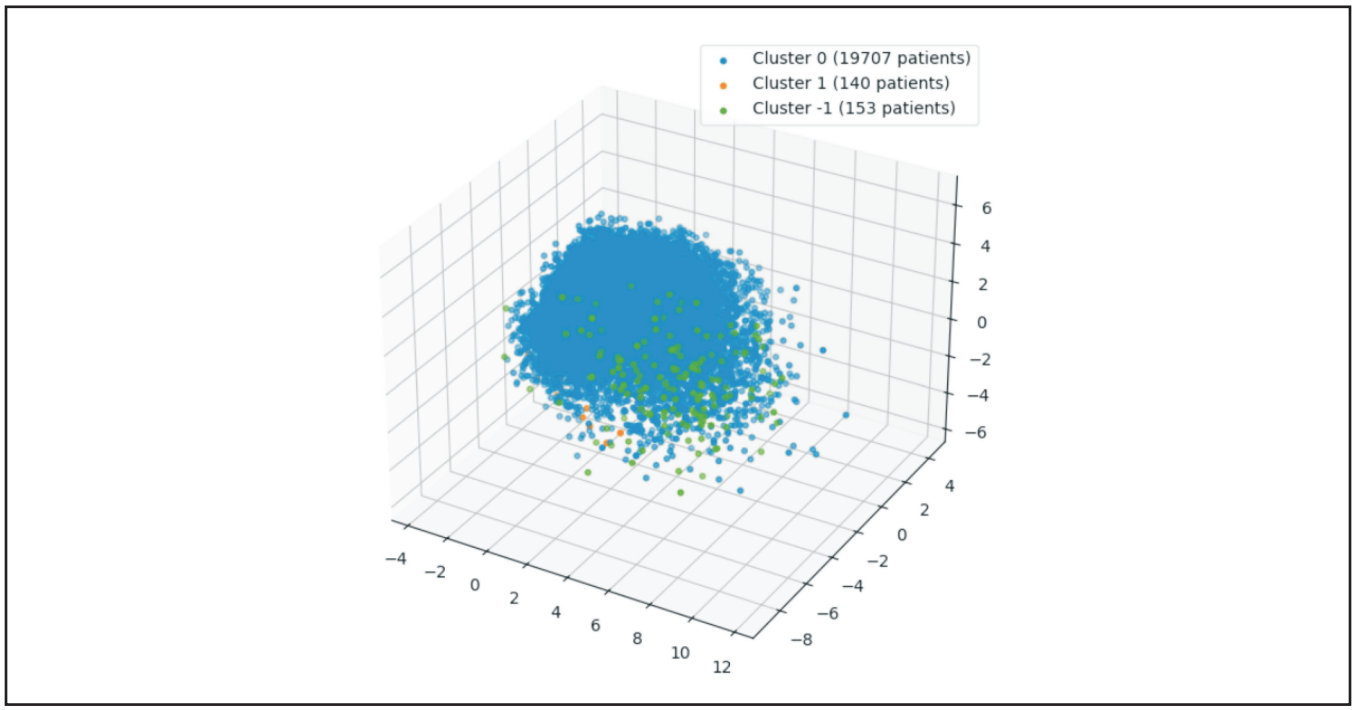


Figure 3: *DBSCAN Clustering of Patient Data*

The 3D visualization of the clustering results shows a highly imbalanced cluster distribution. Cluster 1 and Cluster -1 represent very small and sparse groups. Cluster -1 typically indicates noise or outliers, these data points have no strong association with any main cluster. Cluster 1 represent a minor sub-population with distinct characteristics compared to the main group. Most of the data-points belong to Cluster 0 indicating a dominant and dense patient group that represents the primary population pattern in the dataset. This cluster forms a compact structure in the feature space, suggesting strong similarity among these patients. Overall, the results suggest that the dataset is largely homogeneous with a few isolated or rare patient profiles identified as separate clusters.

4.3 Hierarchical Clustering

Hierarchical Clustering (Agglomerative approach using Ward's method) was employed to the data set comprising patient's information in order to detect groups with similar attributes. The Ward's linkage method is used since it tends to minimize the variances inside clusters, thus allowing for formation of clusters that are more compact and better separated from each other; the silhouette value is calculated for different numbers of clusters k as shown in Table 5.

Table 5: *Silhouette Scores for distict Values of k*

No. of Cluster(k)	Silhouette Score
3	0.3718
4	0.3141
5	0.3034

The clustering algorithm has successfully segregated the dataset into four different clusters, as seen from the Figure 4 on the following page.

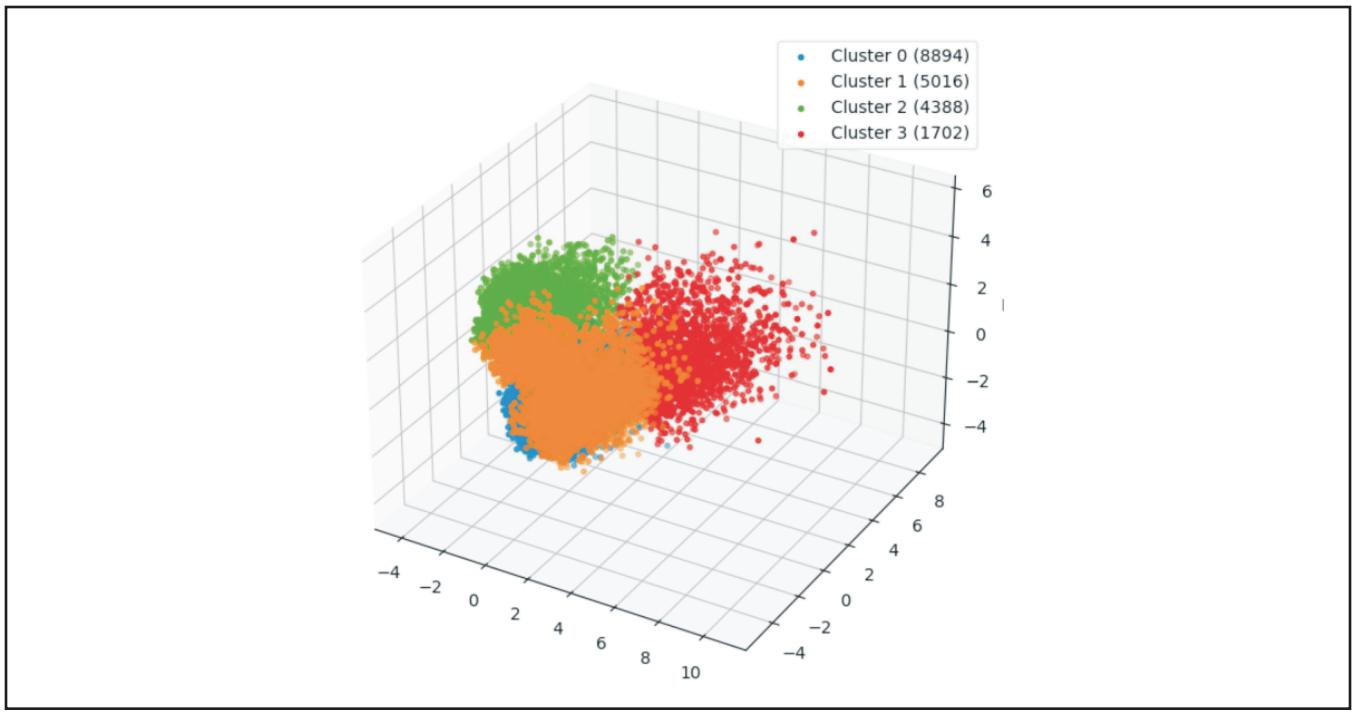


***Figure** 4: Hierarchical Clustering of Patient Data*

The 3D dimensional illustration provides clear insight into how the Ward's method hierarchical clustering algorithm has been successful in generating structured clusters in the transformed data space. Cluster 0 is the largest cluster, implying the dominance of a certain type of pattern in the patient set, whereas Cluster 3 is the smallest and more compact one. Clusters 1 and 2 share some similarities in the sense that they have slight overlaps in some areas. A dendrogram generated using Ward's method, which depicts the hierarchy of the clustering system in the patient data set. In the given diagram, a vertical bar corresponds to two merging clusters, and the linkage height is the measure of the distance between those merging clusters. Ward's clustering algorithm aims at reducing intra-cluster variance. Thus, making the clusters more homogeneous as shown in Figure 5.

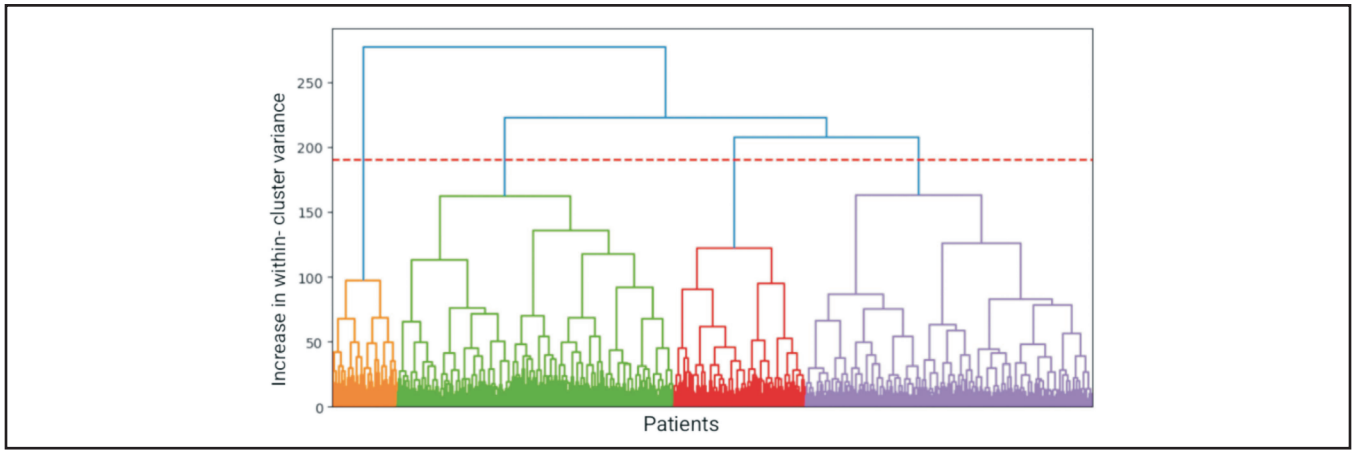


***Figure 5:** Dendogram*

There is also a horizontal cut observed at a linkage distance of about 190, which is represented using a dotted red line. This will result in the formation of four separate clusters. It is therefore implied that the optimum value of k for this set of data is four. The clusters formed here can easily be differentiated due to their large vertical separations. Moreover, the branches on the lower level indicate that data-points within each cluster merge at shorter distances, signifying that there is high similarity among patients who belong to the same cluster. The sharp distinction between major branches indicates that the clustering model is significant and that Ward's method has been successful in capturing the underlying patterns.

5. Comparative Analysis

A detailed comparative analysis between the three clustering techniques: K-Means, DBSCAN, and Hierarchical Clustering was done to assess their efficiency in analyzing the provided data set. The analysis was based on four crucial parameters of

comparison, namely the Silhouette Score, Scalability, Robustness, and Interpretability. The selected parameters will allow obtaining a holistic view of each algorithm from both the performance and practical aspects. The results of such an analysis can be effectively used to choose the best clustering technique for the purpose of the analysis.

5.1 Analysis based on Silhouette Score

The clustering performance has been assessed based on the Silhouette Score, which determines the level of separation and cohesion between clusters.

From Table 6 it can be observed that the highest silhouette score of 0.396 has been acquired by K-Means clustering method, implying high levels of separation between clusters. This can be attributed to the K-means clustering technique, which aims at minimizing distances within the clusters resulting in better cohesion among members. Another successful clustering method was the hierarchical clustering algorithm, which resulted in silhouette score of 0.371.

Table 6: *Silhouette Score of Clustering Algorithms*

Algorithm	SilhouetteScore	Cluster Quality
K-Means	0.396	Very Good
DBSCAN	0.308	Bad
Hierarchical	0.371	Good

However, its performance is quite good because of its capability in dealing with noise and outliers in the data set. Even though it has a lower silhouette score compared to others, it is able to generate useful information through differentiating between dense areas and sparsely populated areas.

In general, from the above results, we can say that the K-Means algorithm produces the most optimal clusters since it performs better with respect to silhouette evaluation. The reason is that it has more cluster cohesion and separation than any other cluster method, indicating that the K-Means algorithm works well with our patient dataset.

5.2 Analysis based on Scalability

The scalability metric is calculated using the time of execution and efficiency of each clustering algorithm in relation to the varying sizes of the input data sets. This assessment makes it easier to gauge the performance capabilities of each algorithm as the amount of data increases, considering their speeds and efficiencies among others. It is only through the scalability aspect that one can establish which algorithm is most appropriate when dealing with large data sets. This is evidenced by the results shown in Table 7.

Table 7: *Scalability Comparison of Clustering Algorithms*

Algorithm	Relative Score	Time Taken
K-Means	9	Less
DBSCAN	6	Average
Hierarchical	5	More

As shown in Table 7 K-means has high scalability as it has a lower computation complexity and converges faster than other clustering methods. On the contrary, DBSCAN exhibits medium-level scalability owing to the fact that its performance relies on the process of neighborhood searches, which increases the computational complexity with an increase in data volume. Hierarchical clustering exhibits low scalability because it requires the computation of pairwise distances and the construction of a dendrogram, leading to high computational cost.

Therefore, K-Means is well-suited for handling large datasets, while Hierarchical clustering is less efficient when the data size increases.

5.3 Analysis based on Robustness

Robustness is computed by measuring the variation in silhouette score before and after introducing noise.

As shown in Table 8 DBSCAN demonstrates high robustness as it can effectively identify and separate noise points, preventing them from influencing cluster formation. However, K-Means is vulnerable due to its high sensitivity towards outliers. These outliers tend to affect the centroids of clusters and thereby affect the process of clustering. Hierarchical method provides moderate robustness, as it can capture data structure to some extent but may still be influenced by noise during the clustering process.

Table 8: *Robustness Score of Clustering Algorithms*

Algorithm	S (Original)	S'(Noisy)	$\lvert S - S' \rvert$	Robustness
DBSCAN	0.308	0.234	0.074	0.926
Hierarchical	0.371	0.134	0.237	0.763
K-Means	0.396	0.024	0.372	0.628

Overall, DBSCAN is the most robust algorithm, while K-Means algorithm is most affected by noise and outliers.

5.4 Analysis based on Interpretability

Interpretability is evaluated based on cluster separability and the ease of understanding the clustering structure.

Feature Simplicity Measure (F_s): The Feature Simplicity score evaluates the simplicity of the clustering algorithm based on the number of required parameters and its structural complexity. It is defined as:

$$F_s = \frac{1}{N_p + N_s} \tag{13}$$

Where:

- N_p = number of parameters required by the clustering algorithm,
- N_s = structural complexity factor of the algorithm.

A lower number of parameters and simpler structure result in a higher F_s value, indicating better feature simplicity.

For the clustering algorithms used in this study:

- **K-Means:** $N_p = 1, N_s = 0.5$

$$F_s = \frac{1}{1+0.5} = 0.66 \tag{14}$$

- **DBSCAN:** $N_p = 2, N_s = 0.5$

$$F_s = \frac{1}{2+0.5} = 0.40 \tag{15}$$

- **Hierarchical Clustering:** $N_p = 1, N_s = 0$

$$F_s = \frac{1}{1+0} = 1.0 \approx 0.99 \tag{16}$$

Table 9: *Interpretability Score of Clustering Algorithms*

Algorithm	C_s (Silhouette)	F_s (Feature Simplicity)	Interpretability
Hierarchical	0.371	0.99	0.675
K-Means	0.396	0.66	0.498
DBSCAN	0.308	0.40	0.354

From Table 9 Hierarchical clustering demonstrates high interpretability as it provides a dendrogram, which visually represents the data point-cluster associations at various degrees of resolution. It helps to understand the relationships better and gives a more comprehensive view of the entire clustering process. K-Means offers moderate interpretability, as the clusters are clearly defined by centroids and can be easily visualized. However, it does not provide insights into hierarchical relationships among data points. In contrast, DBSCAN has low interpretability because it forms clusters based on density without clear boundaries. A high feature simplicity score is assigned to hierarchical clustering due to its intuitive dendrogram representation, which provides a clear and easily interpretable visualization of cluster relationships, especially when using Ward's linkage method. Overall, Hierarchical clustering is the most interpretable method, followed by K-Means, while DBSCAN provides comparatively less intuitive results.

6. Conclusion

This study presents comparative analysis of K-Means, DBSCAN, and Hierarchical clustering techniques, which were applied to analyze a sample of patients. The analysis involved several aspects in order to get an adequate idea of how effective those clustering techniques are. The first interesting aspect of this work is the following: Each of the considered algorithms created its own structure of clusters. K-Means generated 5 clusters, while DBSCAN detected 3 dominant clusters as well as outliers. Finally, 4 significant clusters were detected by Hierarchical clustering approach. This diversity of results does not mean that one algorithm is better than others, on the contrary, it means that these algorithms provide different approaches to the analysis. Patient data are complex and multidimensional, so no approach could catch all peculiarities. The K-Means technique was found to yield the most in-depth classification among the three techniques. Through segmenting the data into five categories, the method helped classify the patients into different health statuses based on their combination of such characteristics as age, weight, BMI, general health status, and diseases. The first group contains 7077 patients that comprised mostly younger people with normal BMI and good health and who rarely suffered from any illness. These patients belong to the low-risk category, and the main aim of preventive measures will be to promote healthy lifestyle. The second cluster contains 5112 patients that included middle-aged people who had an increasing Body Mass Index (BMI) and initial symptoms of diseases such as diabetes or hypertension. Their state was non-critical. However, there were evident signs of potential future problems. Early detection may be considered extremely important in preventing the development of more dangerous diseases. The third cluster has 2059 patients who were primarily older and had prior experience with cardiovascular problems like heart diseases and strokes. Age, bad health condition, and co-morbidities were the main factors that made this clustering happen. It is an important group because their health conditions put them at great risk. Another cluster of 1890 patients has been determined by behavioral aspects, specifically smoking, coupled with conditions such as asthma and COPD. Although these individuals may be in diverse age groups, behavioral factors tend to weigh heavily on their risk profile. For this cluster, preventive measures have to emphasize behavioral changes more than just medical intervention. Cluster five has 5112 patients that has a mixture of moderate characteristics. Those in this category were neither clearly high nor low in risk but rather possessed a combination of moderate body mass index, moderate medical issues, and varied lifestyle risk factors. This demonstrates the concept of mixing in real-life situations where clear separation cannot be achieved and there will always be some overlap between categories. In contrast to other methods, DBSCAN handled the data in a different way using density and not the pre-defined shapes of clusters. Hence, there were 3 major clusters as well as several outlier samples generated by DBSCAN. The first one that contains 19707 patients characterized by the biggest number of samples and was constituted by a dense set of relatively healthy people with feature values similar to each other. The second group contains 140 patients whose symptoms reflected clearly that

they were suffering from a chronic condition. Issues like obesity, diabetes, cardiovascular ailments, and other signs of poor health contributed to the formation of this concentrated cluster. Unlike K-Means clustering where there is fragmentation of the cluster, DBSCAN retains its homogeneity through density. Third cluster contains 153 patients that represented the patients having risks from their lifestyle especially with respect to smokers and respiratory problems. These patients partly matched the patterns found by K-Means clustering but were not clearly visible due to highly dense areas. One of the benefits that the DBSCAN algorithm had over others was the capability to detect outliers. 153 patients are found as outliers. These consisted of patients who exhibited unusual feature combinations, for instance, those who suffered from several severe diseases. Outliers are particularly significant in a medical setting because they represent exceptional cases that deserve urgent and specialized attention. Other algorithms fail to take care of outliers because they assign them to incorrect clusters. Hierarchical clustering analysis gave a unique view of the patients by arranging them in a tree-like pattern. Based on this pattern, 4 major clusters could be distinguished at the right level of separation. The first cluster contains 8894 patients from healthy people with low-risk levels, similar to the first group in the K-Means algorithm. The second cluster contains 5016 that comprised patients with medium-risk levels with rising BMI and developing health problems. In the third cluster, high-risk patients were included 1702 patients who had many chronic diseases and were relatively aged. They were easily identifiable because of their distinct characteristics like age, disease record, and overall health status. Cluster four consists 4388 patients that depended on behavioral aspects such as smoking and respiratory conditions that resulted in patients being classified into a separate group of preventable risk factors.. The unique thing about hierarchical clustering is that it does not restrict the grouping process to a predetermined number of groupings. It demonstrates how small clusters become bigger clusters and gives insight into the relationship between patients in different clusters. This helps us understand how various risk categories relate to each other and the progression between levels. While making a comparison among the three clustering algorithms mentioned above, it can be noted that each of them provides a unique strength. The variance is also shown in the number of clusters formed- five in K-Means, three in DBSCAN, and four in hierarchical clustering, thus revealing that each algorithm has its unique perception regarding similarity. Despite the fact that K-Means is more detailed, DBSCAN is more concerned with denser and more natural groupings, while Hierarchical clustering provides an organized and structured set of clusters. With regard to preventive healthcare, the discovered clusters are directly related to intervention methods. The healthy patients can be treated through health awareness campaigns, and the medium risk ones can undergo intervention methods such as dietary and physical activity changes. High-risk patients should receive constant monitoring services, while patients from behavioral clusters should engage in behavioral intervention methods such as anti-smoking programs.

The clustering approaches used in this study have various applications within the health sector. First, clustering can be applied in patient segmentation through grouping patients that share the same clinical attributes. This makes it possible to implement individualized treatment approaches. Clustering can also play an important role in disease subtyping where healthcare providers will be able to distinguish different types of diseases depending on their characteristics, symptoms, past medical records, and risk factors. Additionally, clustering is useful in discovering drugs by grouping those with similar biological properties, thus leading to the discovery of efficient medications. Lastly, the use of clustering can lead to efficient allocation of healthcare resources because it enables identification of areas that need similar resources.

Overall, this study highlights that cluster analysis is not only a computational process but also a practical technique for gaining insights about complex data sets. Through the grouping of large data into useful clusters, patterns become clear. The conclusions prove that it can be very beneficial to use different clustering techniques for gaining a greater decision-making ability regarding patients' health. Analytically comparing various clustering methods reveals that there is no one particular method that could outshine others in terms of its applicability and efficiency. It is important to remember what exactly the aim of analyzing data is as well as to consider the features of the dataset that is going to be analyzed. When combined, these clustering techniques become highly valuable for obtaining useful results regarding the studied data.

Data Availability Statement

The dataset supporting the findings of this study is publicly available at: *https://www.kaggle.com/datasets/tarekmuhammed/patients-data-for-medical-field*

Acknowledgements

The authors wish to extend their heartfelt thanks to the Department of Mathematics of Shaheed Rajguru College of Applied Sciences for Women, University of Delhi, for their valuable guidance, encouragement, and support rendered to us during the conduct of this research.

Conflict of interest statement

The authors declare that there is no conflict of interest concerning the publication of this research paper.

Ethical Declarations

The study does not include any human subjects, animals, or experimentation that needs ethical review. All data and materials used in the study have been collected and analyzed following relevant academic and ethical principles.

References

1. Aggarwal, C. C., & Reddy, C. K. (2018). Educational and software resources for data clustering. *In Data Clustering* (pp. 607-616). Chapman and Hall/CRC. https://doi.org/10.1201/9781315373515

2. Ahmad, A., & Khan, S. S. (2019). Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access*, 7, 31883-31902. https://doi.org/10.1109/ACCESS.2019.2903568
3. Aljibawi, M., Algabri, H. K., & Rasool, . Z. I. . (2025). Adaptive clustering using enhanced DBSCAN: a dynamic approach to optimizing density-based clustering. Statistics, Optimization & Information Computing, 14(4), 1980–1991. https://doi.org/10.19139/soic-2310-5070-2484
4. Aljohani, A. (2024). Optimizing patient stratification in healthcare: A comparative analysis of clustering algorithms for ehr data. *International Journal of Computational Intelligence Systems*, 17(1), 173. https://doi.org/10.1007/s44196-024-00568-8
5. Andreopoulos, B., An, A., Wang, X., & Schroeder, M. (2009). A roadmap of clustering algorithms: finding a match for a biomedical application. *Briefings in bioinformatics, 10* (3), 297-314. https://doi.org/10.1093/bib/bbn058
6. Bradley, J. V. (1978). Robustnessϵ. *British Journal of Mathematical and Statistical Psychology, 31* (2), 144-152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x
7. Celebi, M. E., Kingravi, H. A., & Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert systems with applications, 40* (1), 200-210. https://doi.org/10.1016/j.eswa.2012.07.021
8. Data of Patients for Medical Fieldshttps://www.kaggle.com/datasets/tarekmuhammed/patients-data-for-medical-field.
9. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).
10. Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering applications of artificial intelligence, 110*, 104743. https://doi.org/10.1016/j.engappai.2022.104743
11. Ghazal, T. M. (2021). Performances of k-means clustering algorithm with different distance metrics. *Intelligent Automation & Soft Computing, 30* (2), 735-742. https://doi.org/10.32604/iasc.2021.019067
12. Harmouch, M. (2021). 17 types of similarity and dissimilarity measures used in data science. *Medium*.
13. Hasibuan, A., Kembuan, D. R., Rantung, V. P., & Tinambunan, M. H. (2023). Optimizing K-Means Clustering: A Comparative Study of Optimization Algorithms For Convergence And Efficiency. *Technium, 16*. https://doi.org/10.47577/technium.v16i.9962
14. Huang, L., Shea, A. L., Qian, H., Masurkar, A., Deng, H., & Liu, D. (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. *Journal of biomedical informatics, 99*, 103291. https://doi.org/10.1016/j.jbi.2019.103291
15. Malik, A., & Tuckfield, B. (2019). *Applied unsupervised learning with R: Uncover hidden relationships and patterns with k-means clustering, hierarchical clustering, and PCA*. Packt Publishing Ltd.
16. Neeraj, K. N., & Maurya, V. (2020). A review on machine learning (feature selection, classification and clustering) approaches of big data mining in different area of research. *Journal of critical reviews, 7* (19), 2610-2626. https://doi.org/10.31838/jcr.07.19.322
17. Ogbuabor, G., & Ugwoke, F. N. (2018). Clustering algorithm for a healthcare dataset using silhouette score value. *Int. J. Comput. Sci. Inf. Technol, 10* (2), 27-37.
18. Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: application and trends. *Artificial intelligence review, 56* (7), 6439-6475.
19. Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. D. F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PloS one, 14* (1), e0210236. https://doi.org/10.1371/journal.pone.0210236
20. Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, 8, 80716-80727.
21. Sun, X. H. (2002). Scalability versus execution time in scalable systems. *Journal of Parallel and Distributed Computing, 62* (2), 173-192.
22. Taiwo, K. A., Olatunji, G. I., & Akomolafe, O. O. (2024). Using Clustering to Segment High-Risk Patients for Tailored Interventions. *Gyanshauryam, International Scientific Referred Research Journal, 7* (4), 251-286. https://doi.org/10.32628/GISRRJ
23. Wani, A. A. (2024). Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions. *PeerJ Computer Science, 10*, e2286. https://doi.org/10.7717/peerj-cs.2286

❖ ❖ ❖