International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

GSJ: Volume 14, Issue 6, June 2026, Online: ISSN 2320-9186

www.globalscientificjournal.com

# BENCHMARKING DATA CLUSTERING ALGORITHMS FOR HEALTHCARE DATA ANALYTICS

**Cheta Franklin Ekweozoh, Ibrahim Mustapha, Chukwunwike Patrick Nwokolo, Usman Salisu Argungu, Martins Ebam**

**Corresponding Author Email**: ekweozohcheta@gmail.com

## ABSTRACT

Healthcare data analytics has been a critical and efficient means of ensuring improvements in care, patient outcome, disease management, operational efficiency, and data-driven decisions in healthcare. The research focused on patient risk stratification in healthcare by benchmarking six clustering algorithms, such as K-Means, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Hierarchical Clustering, Gaussian Mixture Model (GMM), Spectral Clustering, and Deep Clustering. The PIMA India diabetes public health dataset available in Kaggle and UCI machine learning were used for the experiments. Some data preprocessing techniques, such as removing unwanted features, handling missing values, and feature scaling, were implemented prior to applying clustering.

These algorithms were evaluated with these metrics: Silhouette Score, Davies-Bouldin index, computational efficiency, scalability, interpretability, and ability to tolerate noise. It has been found that among the conventional clustering algorithms, DBSCAN provided the highest clustering accuracy in terms of a Silhouette Score of 0.39 and Davies-Bouldin index of 0.76, whereas Deep Clustering provided high clustering accuracy with a Silhouette Score of 0.37 and Davies-Bouldin index of 0.91. While K-Means failed to provide the highest clustering accuracy, it performed excellently in terms of computational efficiency, scalability, simplicity, and interpretability. It can be concluded that K-Means continues to be the most practical, fast, and scalable clustering algorithm for

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

healthcare analytics.

**Keywords:** *Patient risk stratification, Clustering algorithms, K-Means, Benchmarking, Diabetes, Accuracy.*

## 1.0 Introduction

The exponential increase in the volume of healthcare data has converted the healthcare sector into a highly data-centric field, demanding high-end techniques for effective analysis to derive knowledge. Modern healthcare systems constantly produce huge amounts of both structured and unstructured data using various means like electronic health records (EHRs), wearables, medical imaging systems, laboratory management systems, and monitoring systems. According to Singh & Singh (2024) good analysis of the datasets helps for effective and better care delivery, while reducing business expenditure.

Data clustering has become one of the key unsupervised machine learning approaches in healthcare analytics. Data clustering refers to grouping data objects that share similarities while making sure that different data objects are assigned to different clusters. Patient segmentation, disease subgrouping, fraud detection in healthcare, detection of disease outbreak, treatment recommendation systems, and patient risk stratification are some healthcare applications of clustering algorithms (Pushpalatha & Durga, 2023).

Healthcare facilities use patient risk stratification to assess patient danger levels through their clinical indicators and demographic data and disease information. Through patient stratification, healthcare organizations can detect their most dangerous patients while they still have time to distribute medical resources and carry out specialized treatments (Taiwo et al., 2024). Few of the clustering algorithms used in healthcare analytics include K-Means, DBSCAN, Hierarchical Clustering, Gaussian Mixture Models, Spectral Clustering, and Deep Clustering algorithms. Choosing the best clustering algorithm in healthcare analysis is difficult, mainly because the data are noisy, high-dimensional, incomplete, and heterogeneous (Preud'homme et al., 2021). Out of all these clustering algorithms, K-Means clustering has emerged as the most common clustering method because of its computational effectiveness, scalability, and excellent performance in structured health care data sets (Rai et al., 2023). Even though several advanced clustering algorithms have been discovered, K-Means continues to be the benchmarking algorithm in healthcare data analytics.

The current research seeks to benchmark different clustering algorithms in healthcare data analytics and explore why K-Means clustering is the most successful approach in identifying patient risks.

### 1.1 Problem Statement

The analysis of massive amounts of complex patient data has become increasingly difficult in the healthcare industry. The

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

conventional methods of patient classification and risk assessment are not only inefficient but also prone to errors. Moreover, these methods cannot handle the increasing amounts of data produced daily in the healthcare industry. Patient risk stratification enables healthcare professionals to identify high-risk patient groups and forecast future disease development while distributing medical resources most effectively. Nevertheless, healthcare data sets have the following features:

- High-dimensional nature
- Presence of missing values
- Presence of noise and outliers
- Patient heterogeneity
- Massive data size

All these features make the task of clustering challenging and also make it difficult to select an appropriate algorithm. While many clustering algorithms have been proposed, it is important to note that no optimal clustering algorithm exists for healthcare analytics (Rodriguez et al., 2019).

While advanced techniques of clustering such as deep clustering and spectral clustering give very accurate clustering results, they consume large amounts of computing power and also require complicated tuning. On the other hand, simple clustering algorithms such as K-Means are fast yet easily interpretable but may not work well when clusters are oddly structured. Thus, it is vital to conduct an extensive benchmarking test to help determine which clustering algorithm is best suited for healthcare analytics.

### 1.2 Aim and Objectives of the Study

The aim of the research is to benchmark six different clustering algorithms: K-Means, DBSCAN, Hierarchical clustering, Gaussian mixture models, Spectral clustering, and Deep clustering. To verify why the K-Means algorithm is the most suitable for patient risk stratification problems in healthcare analytics. The objectives of the study are as follows:

1. Performing the K-Means clustering technique to handle patient risk stratification and contrasting it with other clustering algorithms.
2. Evaluating the importance of clustering techniques in healthcare analytics.
3. Employing the use of standard measures for evaluating the performance of clustering methods.
4. Verifying why the preferred algorithm is the most suitable.

### 1.3 Research Questions

1. How good is the performance of the K-Means clustering method in handling patient risk stratification?
2. Among the six clustering techniques, which is the best fit that gives a good match within interpretability, accuracy, and scalability?
3. Identify the related issues with clustering datasets.
4. What made the K-Means clustering method frequently used in healthcare even with the presence of advanced algorithms?

## 2.0 Literature Review

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

### 2.1 Clustering in Healthcare Analytics

Clustering is now a pretty important technique in healthcare analytics, mainly because there's so much healthcare data coming in all the time from electronic health records, lab systems, wearable devices, and medical imaging technologies. In general, clustering algorithms help healthcare professionals spot unseen patterns and find similarities and linkages inside patient datasets without needing labels set ahead of time. In healthcare environments, clustering gets used a lot for patient risk stratification, disease subtype discovery, and treatment recommendations, plus sometimes healthcare fraud detection and even epidemic surveillance.

In the view of Ezugwu et al. (2022), clustering approaches may be classified as partitioning, hierarchical, density, grid, and model-based algorithms. The clustering process is not easy for datasets of the healthcare sector since these datasets have issues such as missing value, noise data, heterogeneity of variables, and many others (Wani, 2024). The implementation of good clustering approaches will help healthcare organizations categorize their patients based on some clinical factors.

### 2.2 K-Means Clustering

The K-Means clustering algorithm is a partition-based algorithm that works by partitioning or grouping the data points or observations into k clusters. The algorithm does this by reducing variance within each cluster.

For better emphasis, let's explore K-Means objective function as shown in Figure 1.

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - \mu_i||^2$$

Where:

- $C_i$ represents cluster $i$
- $\mu_i$ represents the centroid of cluster $i$

Figure 1. K-Means objective function (Rodriguez et al., 2019; Ezugwu et al., 2022)

K-Means clustering is pretty much one of the most used partition-type clustering algorithms in healthcare analytics, mainly because it's simple, computationally efficient, and scales well. In practice it does this by splitting a dataset into a fixed number of clusters, then each data item ends up in the group whose centroid is closest. The main goal of K-Means, from the big picture view, is to drive down the intra-cluster variance and, at the same time, boost the separation between clusters so the groups feel more distinct and, kind of, more separate from each other.

In healthcare, K-Means is used a lot for patient risk stratification, disease pattern identification, medical image segmentation, and healthcare resource allocation. In the Rai et al. (2023); Tamrakar et al. (2024) studies, it was noticed that K-Means actually groups the patients together by clinical indicators. The method is a good fit for structured healthcare data because it handles huge patient records efficiently, and at the same time it creates clusters that are easier to interpret by clinicians. That makes it useful for clinical decisions and for more tailored treatment planning.

### 2.3 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm that

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012
ISSN 2278-7763

basically clusters data points by looking at the neighborhood density. In practice, it finds groups by linking points that sit closely together, and then it treats isolated observations as noise or outliers. DBSCAN can be handy especially in healthcare analytics because many medical datasets have irregular patient records, noisy measurements, and sometimes abnormal clinical values that show up without warning. Also, unlike K-Means, DBSCAN does not make you to specify the exact number of clusters, which sounds simple but is a real convenience. The method also tends to work well when the patient groups have irregular shapes and when you want to spot unusual disease patterns. That said, DBSCAN is pretty sensitive when it comes to choosing parameters, and it may have trouble with data where the density changes across the space (Abeer Aljohani, 2024).

DBSCAN neighbourhood density function is expressed in Figure 2

$$N_\varepsilon(p) = \{q \in D \mid dist(p,q) \leq \varepsilon\}$$

Where:

- $\varepsilon$ represents the neighborhood radius
- $D$ represents the dataset
- $dist(p,q)$ represents the distance between data points $p$ and $q$

Figure 2. DBSCAN density function (Rodriguez et al., 2019; Wani, 2024)

**2.4 Hierarchical Clustering**

Hierarchical clustering is a clustering approach; it takes data and lines it up in a tree-shaped form, usually shown as a dendrogram. The process itself can run with agglomerative strategies where it joins smaller groups into larger ones. Or divisive strategies, which instead split those bigger groups into smaller, more detailed sets, and though it's a bit opposite. In healthcare analytics, hierarchical clustering is often applied for disease subtype discovery, patient resemblance checking, and also for various genomic investigations.

One of the benefits of this approach is its ability to interpret the results obtained, where healthcare practitioners can observe the relation between patient groups in the form of a dendrogram. Nevertheless, one of the disadvantages of hierarchical clustering is its high computational complexity and inefficiency on big health data (Psychiatry Research, 2023).

The Single linkage distance function is expressed in Figure 3

$$D(A,B) = \min_{a \in A, b \in B} d(a,b)$$

Where:

- $A$ and $B$ represent two clusters
- $d(a,b)$ represents the distance between data points $a$ and $b$

Figure 3. Hierarchical clustering (Rodriguez et al., 2019; Singh & Singh, 2024)

**2.5 Gaussian Mixture Model (GMM)**

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012
ISSN 2278-7763

Gaussian Mixture Model (GMM) is a probabilistic clustering technique that works on the premise that the data points arise out of a combination of Gaussian distribution functions. As against K-Means, where assignment to a cluster is rigid, GMM gives out soft-clustering results through probability of belonging to various clusters. This clustering method is quite good for handling of healthcare data analysis, as there might be overlapping of patients' data along with uncertainties in medical data.

The method is commonly used for disease progression study, medical diagnosis, and how patient risk might evolve. In many cases GMM can represent intricate patterns in healthcare data better than centroid-driven approaches. Though it tends to need more computation and also a careful tuning of its parameters, so it can be a bit awkward when the healthcare datasets become very large (Rodriguez et al., 2019).

The probability density function is expressed in Figure 4:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

Where:

- $\pi_k$ represents the mixing coefficient
- $\mu_k$ represents the mean vector
- $\Sigma_k$ represents the covariance matrix
- $\mathcal{N}(x|\mu_k, \Sigma_k)$ represents the Gaussian distribution

Figure 4. Gaussian Mixture Model (Rodriguez et al., 2019; Ezugwu et al., 2022)

**2.6 Spectral Clustering**

Spectral clustering is a graph-based clustering technique that makes use of the concept of eigenvalue decomposition and similar matrices in order to group data sets. This clustering technique involves transforming data in the field of healthcare into a graph format and determining clusters from the relationships within the data points. Spectral clustering works best with identifying non-linear or complex patients that other clustering techniques find difficult to identify.

In healthcare analytics, it is commonly applied for medical imaging, genomic analysis, and basically disease subtype discovery too. Even though it shows strong clustering performance, spectral clustering is costly in computation since it leans heavily on matrix decomposition operations (Singh & Singh, 2024).

The normalized Laplacian matrix is expressed in Figure 5:

$$L = D^{-1/2}(D - W)D^{-1/2}$$

Where:

- $L$ represents the normalized Laplacian matrix
- $D$ represents the degree matrix
- $W$ represents the adjacency or similarity matrix

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

Figure 5. Spectral clustering (Rodriguez et al., 2019; Singh & Singh, 2024)

### 2.7 Deep Clustering

Deep clustering involves leveraging the power of deep learning along with clustering algorithms to enhance the feature extraction and representation in clusters. Deep clustering involves applying neural networks or, more specifically, autoencoders to extract useful features before applying the clustering algorithms. There has been good success of deep clustering applied to electronic health records, disease prediction models, medical images, and even in identifying subgroups of patients. Deep clustering can be quite effective in cases where there is complex non-linearity between variables within healthcare datasets. Deep clustering, however, needs huge data sizes, high-performance computing facilities, and proper hyperparameter tuning, all of which adds to the complexity of implementation (Ibna Kowsar et al., 2023).

## 3.0 Research Methodology

### 3.1 Research Design

The research employed an experimental quantitative design to assess different clustering algorithms which researchers use for healthcare data analytics to perform patient risk classification through k-means clustering. The research design is centered on the comparison and performance evaluation of K-Means clustering against the other six clustering algorithms (DBSCAN, Hierarchical Clustering, GMM, Spectral Clustering, and Deep Clustering). The methodology would be carried out using several healthcare datasets and evaluation metrics. The design will employ a benchmark technique, fairness, and consistency by ensuring that all the clustering algorithms used the same test conditions. The robust performance and constraints of the clustering algorithms were analyzed and considered with these criteria: scalability, interpretability, standard of the clustering, and processing performance of the algorithm. According to Gagolewski (2022), in cases of clustering research, test benchmarking is commonly utilized for the goal of comparing algorithms over several datasets and metrics.

The methodology workflow consisted of the following stages:

1. Healthcare dataset gathering or collection
2. Data preprocessing and cleaning
3. Feature selection along with normalization
4. Clustering algorithm implementation (or application)
5. Cluster evaluation and benchmarking
6. Comparative performance analysis overall
7. Interpretation of patient risk groups and labelling

### 3.2 Data Source

The datasets used in this research work have been sourced from an openly accessible machine learning and health care repository. Open-source healthcare datasets have been extensively employed in clustering research since they serve as standard datasets for benchmarking clustering algorithms (Javed et al., 2020).

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012
ISSN 2278-7763

The healthcare datasets were sourced from UC Irvine Machine Learning Repository, Kaggle, and Scikit-learn Datasets. The selected datasets comprised numerical data that could be used for clustering patients.

### 3.3 Dataset

PIMA Indian Diabetes healthcare dataset was picked for benchmarking experiments on the grounds of its relevance to patient risk analysis and healthcare analytics. The selected health dataset with its instances, features, and application is shown in Table 1.

*Table 1. Selected Healthcare Dataset*

| Dataset | Instances | Features | Application |
|---|---|---|---|
| Diabetes | 768 | 8 | Diabetes risk stratification |

This dataset consists of the medical records of patient such as number of times pregnant, glucose level, diastolic blood pressure, triceps skin fold thickness, body mass index, insulin level, diabetes pedigree function, age, and class variable. It was used to cluster patients based on their diabetes risks. Figure 6 shows the data or features of the dataset after loading.

```
# Loading dataset
df = pd.read_csv('diabetes.csv')
```

```
df.head()
```

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age | Outcome |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 2 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 3 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 4 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |

Figure 6. Features of the dataset

### 3.4 Data Preprocessing

Before clustering, data preprocessing was done to lift the overall quality and make everything line up across all the datasets. In healthcare, data can be messy, with missing values, unwanted noise, repeated entries, and also inconsistent scales for the features, and that kind of variance can really drag down clustering performance (Preud'homme et al., 2021).

The steps taken included:

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012
ISSN 2278-7763

**1. Treating Missing Values**

Missing values were spotted and handled with mean imputation for numerical variables and mode imputation for categorical variables. Handling missing values prevents skewness.

**2. Data Cleaning**

Duplicate patient records and those inconsistent entries were completely removed. Noise was eliminated from the data through outlier detection techniques.

**3. Feature Selection**

For patient risk stratification, a set of relevant healthcare variables was selected so the clustering quality could be improved and also keep the dimensionality reduced.

**4. Data Normalization**

Features in healthcare datasets usually vary significantly based on numerical scale. Data normalization was therefore performed using Z-score standardization.

Standardization made all the healthcare variables align on the same kind of scales, and it also boosted the clustering results.

**5. Dimensionality Reduction**

Principal Component Analysis (PCA) was used on those high dimensional data sets so the computational workload wouldn't get too heavy, while still keeping the major variance that matters inside the healthcare information.

**3.5 Clustering Algorithms**

These six clustering algorithms were benchmarked as represented in Table 2.

*Table 2. Clustering Algorithms*

| Algorithm | Clustering Type |
|---|---|
| K-Means | Partition-based |
| DBSCAN | Density-based |
| Hierarchical Clustering | Hierarchical |
| Gaussian Mixture Model | Probabilistic |
| Spectral Clustering | Graph-based |
| Deep Clustering | Deep learning-based |

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

### 3.6 Experimental Environment

Python programming language and libraries such as pandas, numpy, matplotlib, and scikit-learn, were used for this test.

Hardware documentation: Microsoft Windows 11 pro, Intel Core i5 processor, 16GB RAM.

Integrated Development Environment (IDE): Jupyter Notebook.

Software programs: Scikit-learn, Python 3.11, Tensorflow, Numpy, Pandas, and Matplotlib.

### 3.7 Evaluation Metrics

Different benchmark metrics were used for evaluation across the six clustering algorithms for general performance review as represented with the roles of each in Table 3.

*Table 3. Roles of Evaluation Metrics*

| Evaluation Metric | Function |
|---|---|
| Silhouette Score | Evaluates coherency and how the Cluster separates |
| Davies–Bouldin Index | Measures how similar the clusters actually are |
| Computational Time | Measures how efficient the algorithm is |
| Scalability | Evaluates the performance on large datasets |
| Noise Handling | Measures how robust it is to outliers |
| Interpretability | How easy it is for physicians to use in practice |

### 3.8 Comparative Benchmark Process

Each clustering algorithm was trained and checked on the chosen healthcare dataset using the same preprocessing conditions, exactly the same. Benchmarking comparisons were done around a few axes, including:

- Clustering quality
- Computational efficiency
- Scalability
- Noise robustness
- Clinical interpretability

Overall, the benchmarking framework helped with a rather objective evaluation of clustering methods for healthcare analytics and also

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012
ISSN 2278-7763

for patient risk stratification uses. It was kind of straightforward but still, you know, dependable in the way it was set up.

## 4.0 Experimental Results, Comparative Results, and Discussion

### 4.1 Experimental Results

The clustering techniques were developed and validated using the chosen healthcare datasets, using the same preprocessing steps. The studies considered the process of risk stratification through clustering of patients into groups depending on similarities in their clinical indicators such as pregnancies, glucose level, blood pressure, skin thickness, BMI, and other healthcare factors.

The benchmarking experiments were used to look at how well K-Means, DBSCAN, Hierarchical Clustering, Gaussian Mixture Model (GMM), Spectral Clustering, and Deep Clustering performed, based on the Silhouette Score, Davies–Bouldin Index, computational time, scalability, interpretability, and how robust the methods were when dealing with noisy healthcare records.

The experiment outcomes showed great differences in the effectiveness of clustering algorithms working with healthcare data. K-Means proved its stability to produce the balanced cluster solutions with good computational speed and cluster interpretability. K-Means effectively segregated the patients into different risk categories depending on their healthcare factors. Three clusters were set; cluster 0 indicates low-risk patients, cluster 1 indicates medium-risk patients, while cluster 2 indicates high-risk patients, as illustrated in Figure 7. It scored 0.2018 silhouette score and 1.7532 davies-bouldin index for the evaluation metrics.

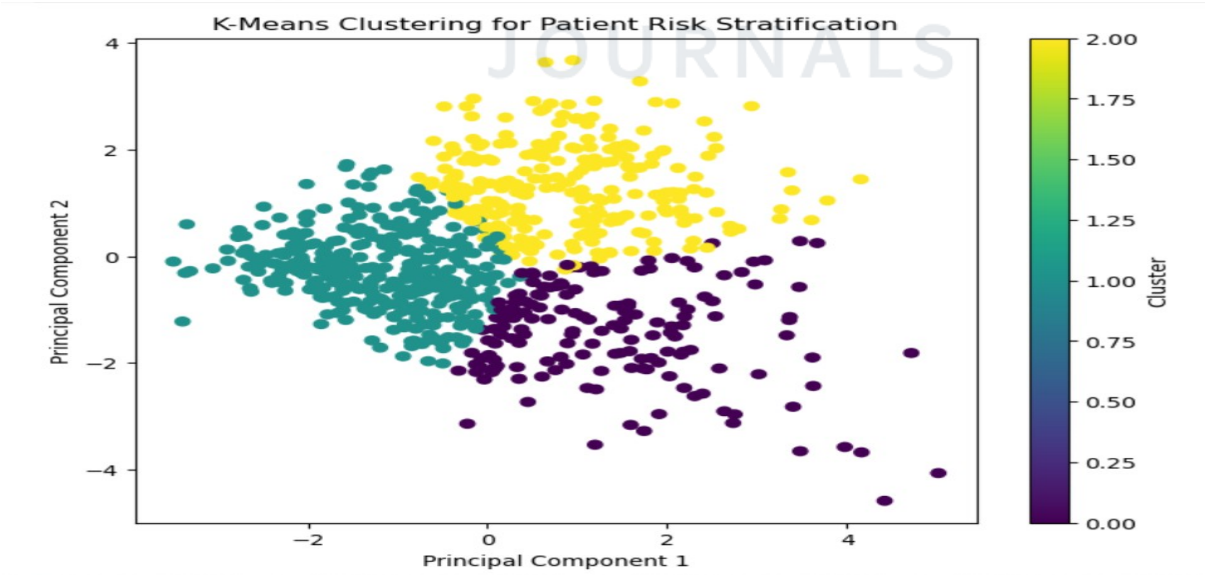


Figure 7. K-Means clustering algorithm

DBSCAN performed satisfactorily in detecting outliers and noisy patient records; however, DBSCAN algorithm had issues dealing with uneven density distribution datasets. Two clusters were set; it had silhouette score 0.3947 and davis-bouldin index 0.7559. The data visualization of the cluster algorithm for patient risk stratification is shown in Figure 8.

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

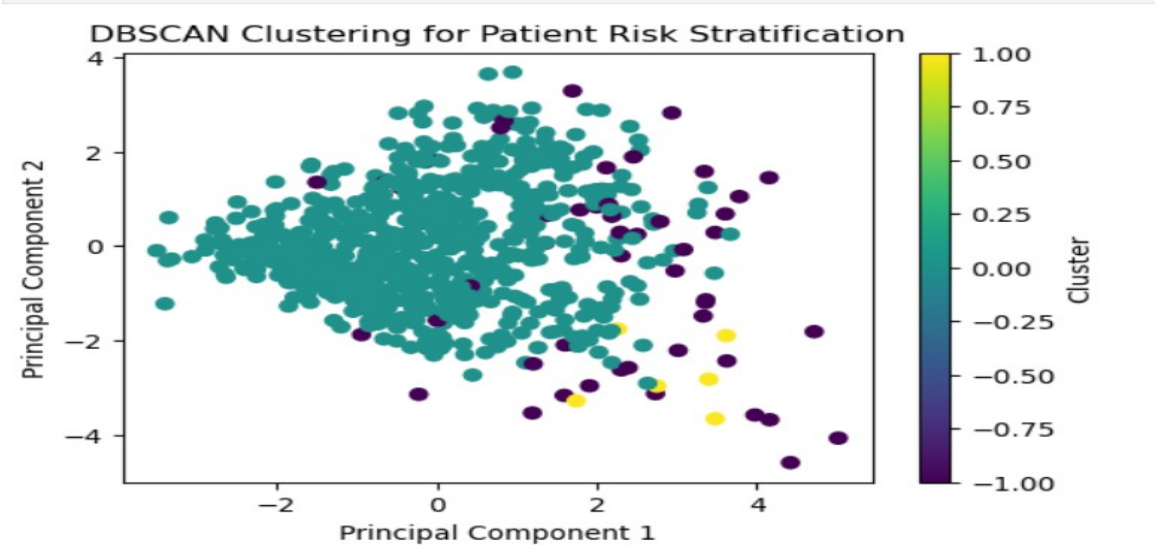


Figure 8. DBSCAN clustering algorithm

Hierarchical clustering yielded highly interpretable dendrograms, although the computation became more complex for large health datasets. On performance, it scored 0.1953 and 1.8497 as silhouette score and davis-bouldin index respectively. The number of clusters used is 3 and data visualization for patient risk stratification is shown in Figure 9.

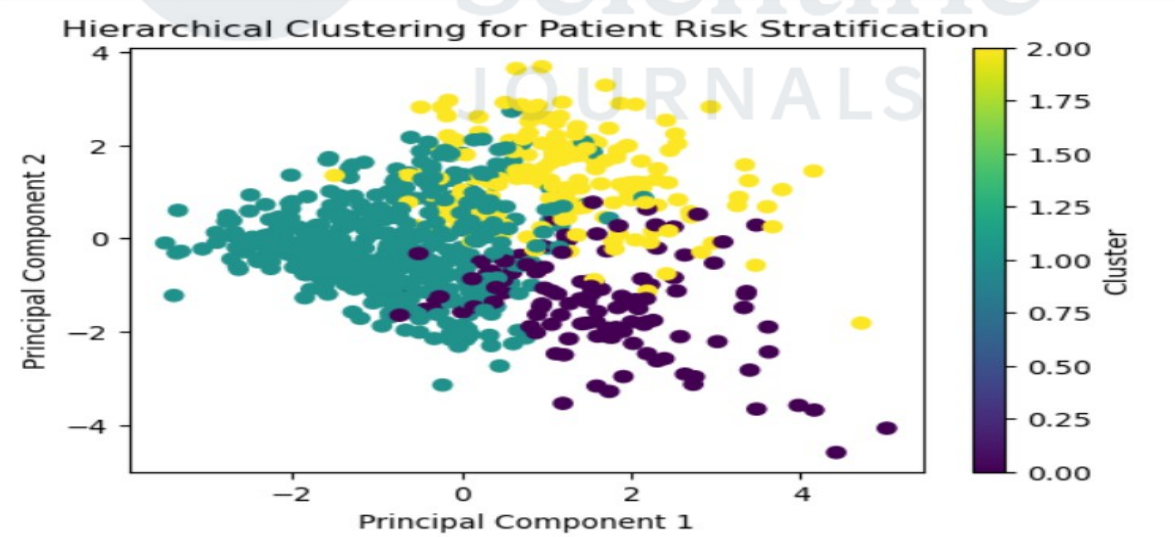


Figure 9. Hierarchical clustering algorithm

The Gaussian Mixture Model showed decent probabilistic clustering results, but it needed more compute resources and careful parameter tuning too. It had silhouette score 0.0256 and davies-bouldin index 2.8319. Three clusters were set and the visualization plot is shown in Figure 10.

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

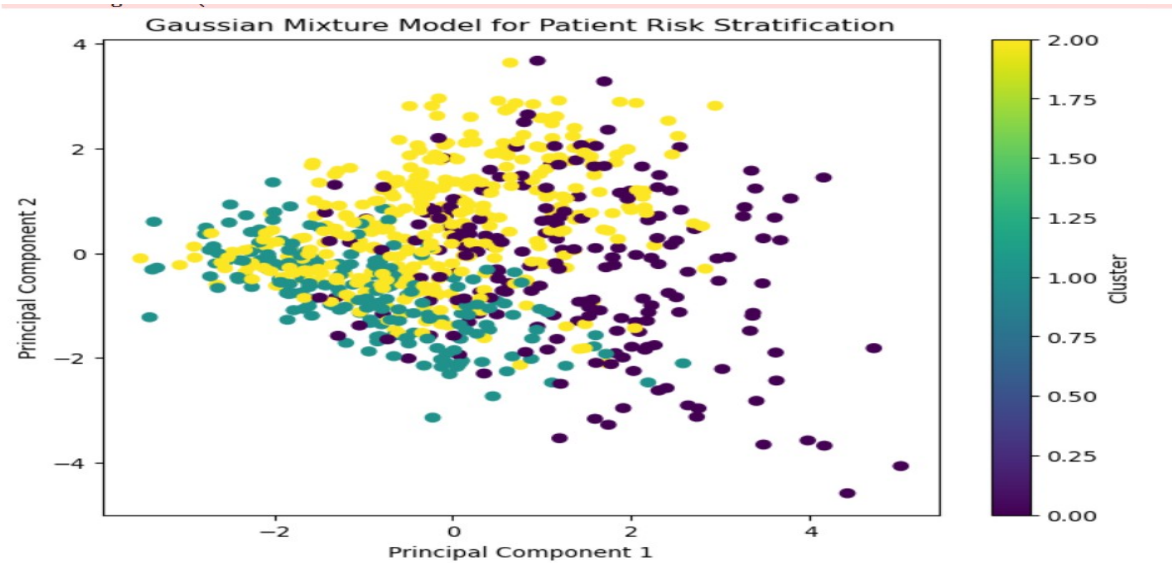


Figure 10. Gaussian Mixture Model clustering algorithm

Spectral clustering got a solid clustering quality on difficult healthcare relationships, yet it didn't scale well when datasets got big, mainly because there were heavy matrix decompositions. Three clusters were used, had silhouette score 0.1611, davis-bouldin index 1.8759, and visualization plot for patient risk stratification shown in Figure 11.

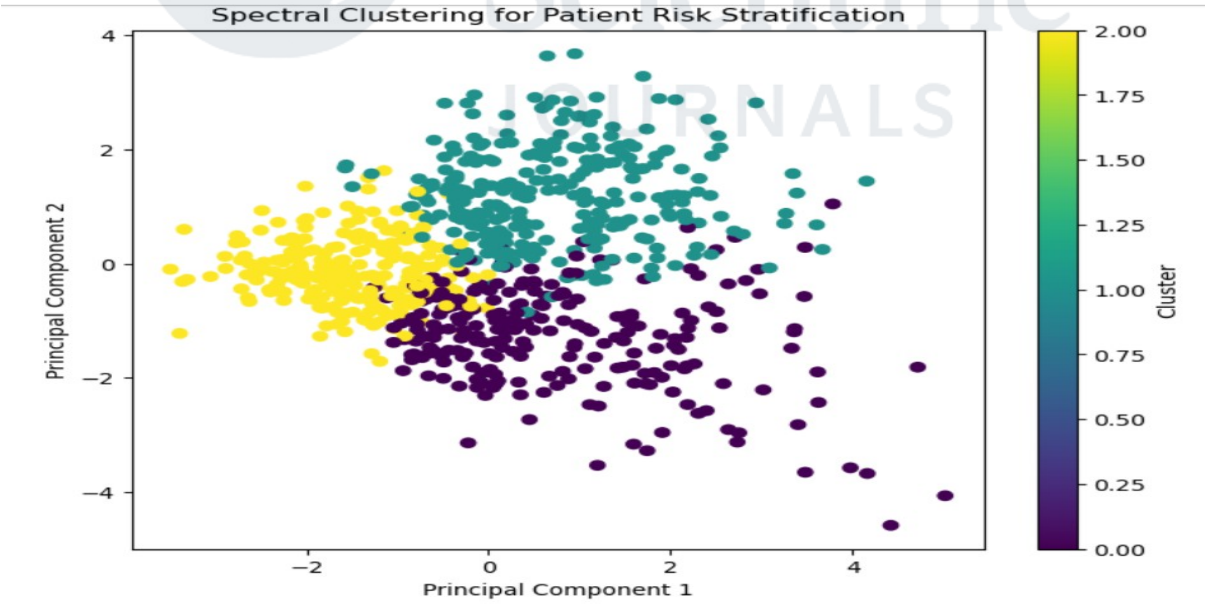


Figure 11. Spectral clustering algorithm

Deep Clustering provided good results regarding clustering accuracy and representation quality, especially in high-dimensional healthcare data. Nevertheless, the approach required substantial computational capabilities, training set size, and longer processing time. Got Silhouette score 0.3742, davis-bouldin index 0.9127 and 3 clusters. The data visualization plot for patient risk stratification is shown in Figure 12.

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

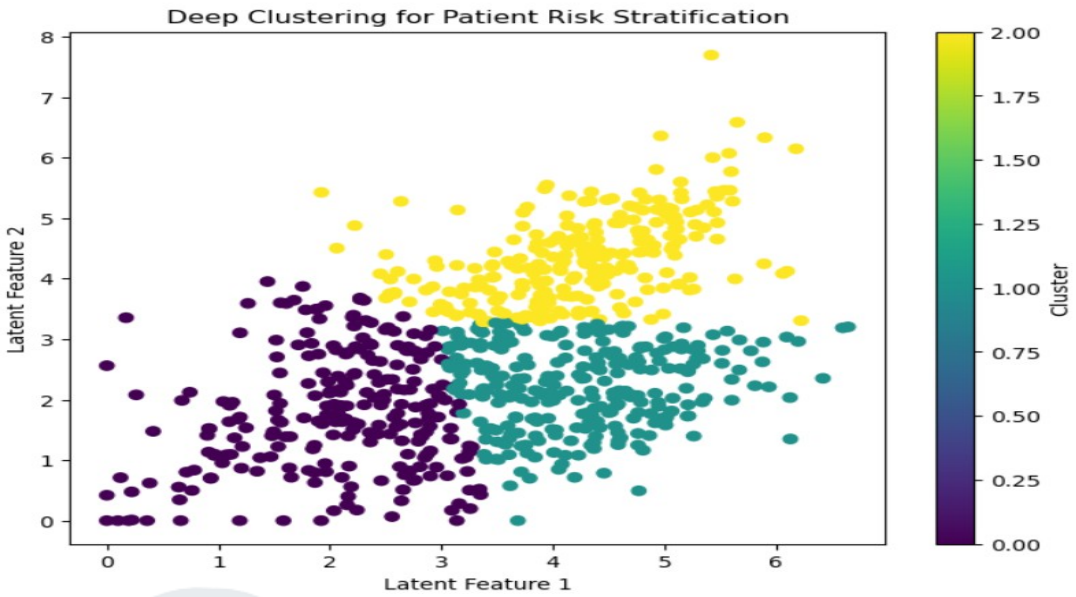


Figure 12. Deep clustering algorithm

K-Means performed well across all parameters with a satisfactory patient grouping and moderate clustering quality.

### 4.2 Comparative Results

Table 4 shows/displays relative performances and comparative results of clustering algorithms.

*Table 4. Comparative performance*

| **Algorithm** | **Silhouette Score** | **Davies-Bouldin Index** | **Computational Speed** | **Scalability** | **Interpretability** | **Noise Handling** |
|---|---|---|---|---|---|---|
| K-Means | 0.20 | 1.75 | Excellent | Excellent | Excellent | Moderate |
| DBSCAN | 0.39 | 0.76 | Moderate | Moderate | Moderate | Excellent |
| Hierarchical clustering | 0.20 | 1.85 | Slow | Poor | Excellent | Moderate |
| Gaussian Mixture Model | 0.03 | 2.83 | Moderate | Moderate | Good | Moderate |
| Spectral clustering | 0.16 | 1.88 | Slow | Poor | Moderate | Good |
| Deep clustering | 0.37 | 0.91 | Very slow | Poor | Moderate | Excellent |

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

The comparative analysis from Table 4 shows that DBSCAN and Deep clustering numerically outperformed the other clustering algorithms based on the criteria of higher silhouette score and lower davis-bouldin index.

The good performance DBSCAN clustering is attributed to the fact that it can effectively outliers and noisy dataset, acknowledging that diabetes risk groups overlap a lot in practice. DBSCAN is built for finding density-based structures, not just neat clusters, it tends to separate those thicker and thinner regions.

Deep clustering performed good as well because it captures latent representations, reduces noise from features, and provides better cluster separation. But it depends on tensorflow, requires GPU power, takes longer time to train, and often requires tunning. Though it lacks practicality in healthcare particularly in Africa due to low-resource hospitals and developing healthcare systems.

However, K-Means clustering is the preferred algorithm due to these factors;

- Computational speed – faster and more efficient.
- Scalability – manages large datasets effectively.
- Interpretability – easy to comprehend and interpret by medical practitioners
- Less complexity – requires minimal parameter settings and runs on resource-limited environment.
- Deployment – readily and easier hospital deployment.

Table 5 outlines the advantages and restrictions of clustering algorithms

*Table 5. Pros and Restrictions of Clustering Algorithms*

| **Algorithm** | **Pros** | **Restrictions** |
|---|---|---|
| K-Means | Speed, Scalability, and Simplicity | Dependent on initial centroids selection |
| DBSCAN | Noise tolerance and outlier management | Dependence on parameters |
| Hierarchical clustering | Dendogram representation | Computational complexity |
| Gaussian Mixture Model | Probabilistic clustering | Resource intensive parameter estimation |
| Spectral clustering | Non-linear structures handling | Lack of scalability |
| Deep clustering | Complex representations learning | Resource intensive computation |

**4.3 Discussion**

The results of this study indicated that there were some significant variations in the effectiveness of clustering algorithms for

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012
ISSN 2278-7763

clustering health care data for predicting risks. DBSCAN performed better among the other clustering techniques because of its capability of detecting dense clusters and dealing with noise in health care data, which gave it a Silhouette Score of 0.39 and Davies-Bouldin Index of 0.76.

Deep Clustering algorithm was highly effective with a Silhouette Score of 0.37 and Davies–Bouldin Index of 0.91 but needed more computational costs and higher complexities than traditional techniques.

K-Means clustering got a Silhouette Score of 0.20 and the Davies–Bouldin Index of 1.75. Even if its clustering accuracy were lower than DBSCAN and Deep Clustering, it still had very good computational speed, a nice level of scalability, and plain interpretability. In a way it feels simple, almost straightforward, and that helps it work in practical healthcare scenarios, especially when resources are tight.

Both algorithms, Hierarchical Clustering and Spectral Clustering had a fair performance, though they encountered scalability and computational constraints. Gaussian Mixture Model performed poorly due to overlap of the data.

Generally, the research showed that cluster accuracy was achieved by DBSCAN and Deep clustering algorithms. However, K-Means clustering algorithm is preferred as a more practical, faster and scalable choice for the purpose of patient risk stratification in healthcare.

## 5.0 Conclusion, Recommendation, and Further Work

### 5.1 Conclusion

In this research work, the performance of six clustering techniques in healthcare data analysis with a special focus on patient risk stratification was benchmarked against a healthcare dataset. These clustering methods included K-Means, DBSCAN, Hierarchical Clustering, Gaussian Mixture Model, Spectral Clustering, and Deep Clustering.

From the findings of the experiment, it can be deduced that DBSCAN and Deep Clustering provided better clustering performance in terms of Silhouette Score and Davies-Bouldin Index. DBSCAN managed to deal well with noisy health care data sets, while Deep Clustering enhanced the representation of the features of the data through autoencoder.

While K-Means clustering performed poorly in terms of clustering accuracy when compared with DBSCAN and Deep Clustering, K-Means produced high computational efficiency, scalability, interpretability, and easy implementation. Therefore, K-Means clustering can be regarded as very useful in healthcare analytics due to its suitability for healthcare settings with poor computational resources.

In conclusion, the analysis above proves that no algorithm for clustering is perfect across all applications. Although clustering algorithms such as DBSCAN and Deep Clustering perform better in clustering, K-Means can be considered the best clustering algorithm that can be used in patient risk stratification within the healthcare industry.

### 5.2 Recommendation

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012
ISSN 2278-7763

The recommendations arising from the research findings include the following:

1. Healthcare organizations are advised to utilize K-Means clustering in risk stratification of their patients due to its scalability, efficiency, and clustering superiority.

2. Healthcare organizations are recommended to incorporate the use of clustering analytics in EHRs to inform healthcare delivery and decision-making.

3. Researchers are encouraged to develop hybrid clustering models through the combination of clustering techniques such as K-Means and other optimization algorithms.

4. Prior to the implementation of any clustering model, healthcare datasets must be preprocessed to ensure normalization, feature selection, and handling of missing values.

5. Health authorities are encouraged to adopt the use of healthcare analytics to enhance surveillance of diseases, healthcare delivery planning, and patient management.

6. The African healthcare system should develop infrastructures for healthcare data analytics to optimize healthcare delivery processes using the available healthcare resources.

## References

1.Gagolewski, M. (2022). A framework for benchmarking clustering algorithms. *SoftwareX*, *20*, 101270. https://doi.org/10.1016/j.softx.2022.101270

2. Javed, A., Lee, B. S., & Rizzo, D. M. (2020). A benchmark study on time series clustering. *Machine Learning with Applications*, *1*, 100001. https://doi.org/10.1016/j.mlwa.2020.100001

3. Singh, J., & Singh, D. (2024). A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery:

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

Taxonomy, challenges, applications and future prospects. *Advanced Engineering Informatics*, *62*, 102799. https://doi.org/10.1016/j.aei.2024.102799

4. An overview of clustering methods with guidelines for application in mental health research. (2023). *Psychiatry Research*, *327*, 115265. https://doi.org/10.1016/j.psychres.2023.115265

5. Jain, K., Pisharody, U., Singh, M., & Sinha, B. B. (2026). Benchmarking clustering techniques: insights for comparative analysis and algorithm selection: a survey. *Knowledge and Information Systems*, *68*(1). https://doi.org/10.1007/s10115-026-02709-1

6. Preud'homme, G., Duarte, K., Dalleau, K., Lacomblez, C., Bresso, E., Smaïl-Tabbone, M., Couceiro, M., Devignes, M.-D., Kobayashi, M., Huttin, O., Ferreira, J. P., Zannad, F., Rossignol, P., & Girerd, N. (2021). Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-83340-8

7. Shand, C., Allmendinger, R., Handl, J., Webb, A., & Keane, J. (2021). HAWKS: Evolving Challenging Benchmark Sets for Cluster Analysis. *IEEE Transactions on Evolutionary Computation*, *26*(6), 1206–1220. https://doi.org/10.1109/tevc.2021.3137369

8. Wani, A. A. (2024). Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions. *PeerJ Computer Science*, *10*, e2286–e2286. https://doi.org/10.7717/peerj-cs.2286

9. Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLOS ONE*, *14*(1), e0210236. https://doi.org/10.1371/journal.pone.0210236

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. Pushpalatha, M. L., & Durga, R. (2023). *Review on Various Clustering Algorithm on Healthcare based Cloud Analytics*. https://doi.org/10.1109/icssit55814.2023.10061079

12. Ibna Kowsar, Rabbani, S. B., Kazi, & Samad, M. D. (2023). Deep Clustering of Electronic Health Records Tabular Data for Clinical Interpretation. *PubMed*. https://doi.org/10.1109/ictp60248.2023.10490723

13. Yang, W.-C., Lai, J.-P., Liu, Y.-H., Lin, Y.-L., Hou, H.-P., & Pai, P.-F. (2024). Using Medical Data and Clustering Techniques for a Smart Healthcare System. *Electronics*, *13*(1), 140. https://doi.org/10.3390/electronics13010140

14. Al-Khamees, H. A. A., Al-Slivani, M. M., Kadhim, M. S., Radhi, A. D., Sani, N. S., Al-Amri, R. M., Wahit, F., & Afira Sani, M. A. (2026). Enhancing classification accuracy in medical datasets using a hybrid distance and cluster refinement-based K-means clustering method. *Scientific Reports*, *16*(1). https://doi.org/10.1038/s41598-025-30176-1

15. Abeer Aljohani. (2024). Optimizing Patient Stratification in Healthcare: A Comparative Analysis of Clustering Algorithms for EHR Data. *International Journal of Computational Intelligence Systems*, *17*(1). https://doi.org/10.1007/s44196-024-00568-8

International Journal of Advancements in Research & Technology, Volume 1, Issue 5, October-2012

ISSN 2278-7763

16. Rai, A. K., Upendra Singh Aswal, V. Saravanan, N SHALINI, Dwivedi, S. P., & Kumar, N. (2023). *Patient Clustering Optimization With K-Means In Healthcare Data Analysis*. 1–7. https://doi.org/10.1109/icaiihi57871.2023.10489428

17. Prajapati, S., & Kumar Timalsina, A. (n.d.). *Proceedings of 12 th IOE Graduate Conference Risk Stratification in Healthcare Data using Clustering Algorithms*. https://conference.ioe.edu.np/publications/ioegc12/IOEGC-12-206-12302.pdf

18. Taiwo, K. A., Olatunji, G. I., & Opeoluwa Oluwanifemi Akomolafe. (2024). Using Clustering to Segment High-Risk Patients for Tailored Interventions. *Gyanshauryam International Scientific Refereed Research Journal*, *7*(4), 251–286. https://www.researchgate.net/publication/394656410_Using_Clustering_to_Segment_High-Risk_Patients_for_Tailored_Interventions

19. Coombes, C. E., Liu, X., Abrams, Z. B., Coombes, K. R., & Brock, G. (2021). Simulation-derived best practices for clustering clinical data. *Journal of Biomedical Informatics*, *118*, 103788. https://doi.org/10.1016/j.jbi.2021.103788

20. Robinson, G., Peng, J., Dönnes, P., Coelewij, L., Naja, M., Radziszewska, A., Wincup, C., Peckham, H., Isenberg, D. A., Ioannou, Y., Pineda-Torra, I., Coziana Ciurtin, & Jury, E. C. (2020). *Disease-associated and patient-specific immune cell signatures in juvenile-onset systemic lupus erythematosus: patient stratification using a machine-learning approach*. *2*(8), e485–e496. https://doi.org/10.1016/s2665-9913(20)30168-5

21. Tamrakar, P., Pathak, G. R., Lal, M., Goel, A., & Bhende, M. (2024). Patient Clustering Optimization With K-Means in Healthcare Data Analysis. *2024 International Conference on Recent Innovation in Smart and Sustainable Technology (ICRISST)*, 1–6. https://doi.org/10.1109/icrisst59181.2024.10921818.