Journal of Technology and System Information Vol: 3, No 1, 2026, Page: 1-38

Clustering Performance on Heart Disease Data: Effects of Distance Metrics and Scaling

Ibrahim Akbas*, Yavuz Selim Taspinar, Murat Koklu

Selçuk University

DOI:
https://doi.org/10.47134/jtsi.v3i1.5336
*Correspondence: Ibrahim Akbas
Email: ibrahimakbas4242@gmail.com

Received: 22-11-2025
Accepted: 22-12-2025
Published: 22-01-2026

Copyright: © 2026 by the authors. Submitted for open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Abstract: Cardiovascular diseases (CVD) are one of the leading causes of morbidity and mortality worldwide, requiring advanced analytical approaches to identify early-stage risk groups and classify patient profiles in greater detail. The aim of this study is to reveal latent patient subgroups associated with CVD using unsupervised machine learning methods on clinical data. In this context, a dataset consisting of 11 clinical variables from 303 patients who visited the VA Medical Center in Long Beach, California, was analyzed. During the preprocessing stage, missing observations were eliminated, only numerical variables were used, and both z-score standardization and min–max normalization were applied to the data. Subsequently, hierarchical clustering analyses were performed using single, complete, and average linkage approaches based on Euclidean and cosine distance measures) (the number of possible clusters for different distance–scaling combinations was evaluated using the Elbow and Silhouette measures. The results obtained showed that the 4-cluster solution, particularly under the complete and average linkage methods, represented the data structure in the most clinically explanatory manner. The similarity between the clustering results obtained using the k-means algorithm with Euclidean distance in standardized data and cosine distance in normalized data was calculated as the Rand Index (RI) = 0.8179) (this value demonstrated that the cluster structure was largely preserved despite different distance metrics and scaling strategies. The findings demonstrate that unsupervised learning-based clustering approaches provide a useful tool for defining meaningful risk classes within heterogeneous patient populations based on clinical datasets and for conducting comparative clinical evaluations between these classes.

Keywords: Cardiovascular Diseases, Clinical Data Analysis, Unsupervised Machine Learning

Introduction

Cardiovascular diseases (CVD) are recognized as one of the most common and serious public health problems in both developed and developing countries today. While less than 10% of deaths worldwide were attributed to CVD in the early 20th century, this rate had risen to 30% by 2005, demonstrating a marked increase in the burden of cardiovascular disease over the years. It is reported that approximately 80% of the global cardiovascular mortality burden occurs in low- and middle-income countries (Gaziano et al, 2006). World Health Organization data clearly show that heart and vascular diseases are one of the main determinants of global mortality rates and pose a serious public health threat (Erdem et al, 2024) (Erdem et al, 2023). CVD, including stroke, is among the leading causes of morbidity and mortality in the United States) (epidemiological estimates indicate that approximately 62 million individuals have cardiovascular disease and 50 million individuals have

hypertension. In 2000 alone, more than 946,000 deaths were attributed to CVD, accounting for approximately 39% of all deaths (Nabel, 2003). Similarly, in China, cardiovascular diseases are the leading cause of both mortality and premature death, accounting for approximately 40% of all deaths. Globally, China and India are among the countries with the highest burden of cardiovascular disease) (in India, CVD accounts for approximately one-quarter of all deaths, making it the leading cause of death. Among these deaths, ischemic heart disease and stroke account for more than 80% of CVD-related mortality (Prabhakaran et al, 2016) (Zhao et al, 2019).

In this context, early detection of heart disease (Capotosto et al, 2018) and reliable identification of key risk factors are critically important for both improving individual quality of life and reducing the economic burden on the healthcare system. Traditional diagnostic methods rely on clinical procedures such as electrocardiography, exercise testing, echocardiography, and invasive angiography (Gorenoi et al, 2012) (Popp, 1976) (Shilaskar & Ghatol, 2013). However, the disadvantages of these methods, such as high cost, time requirements, dependence on specialists, and limited accessibility in some cases, increase the need for complementary and more scalable approaches. At this point, artificial intelligence (AI) and machine learning (ML)-based methods, influenced by technological developments, have emerged as an important research area for the prediction, risk classification, and early diagnosis of heart disease (Krittanawong et al, 2020).

Machine learning is a powerful set of algorithms that enables meaningful patterns to be extracted from large and complex data sets. Within this framework, different types of algorithms are widely used, such as supervised (classification and regression), unsupervised (clustering and dimension reduction), semi-supervised, and reinforcement learning subfields, as well as tree-based methods, support vector machines, ensemble approaches, and artificial neural networks/deep learning architectures. These algorithms are also widely adopted in various industrial and service sectors, in addition to the healthcare field, to improve operational processes, strengthen decision support mechanisms, and optimize resource planning (Hayta et al, 2023) (Koklu & Sabancı, 2016) (Yasin & Koklu, 2025). In recent years, the use of AI and ML techniques in healthcare has increased significantly) (they have become an important decision support tool in various clinical problems such as predicting emergencies, analyzing disease populations, classifying clinical conditions, predicting immune responses, and detecting and classifying numerous disease classes related to different organ systems (Alanazi, 2022) (Cinar et al, 2022) (Saritas et al, 2025) (Taspinar et al, 2022) (Taspinar et al, 2024). However, the practical applicability and clinical interpretability of findings obtained from ML-based models in healthcare services are still debatable) (nevertheless, the process of integrating these methods into healthcare systems is accelerating significantly (Habehh & Gohel, 2021). Through machine learning techniques, parameters such as patient age, gender, blood pressure, cholesterol and blood sugar levels, and obesity can be evaluated together to predict the likelihood of developing hypertension and CVD with high accuracy (Shrivastava et al, 2023).

Studies in the literature show that different ML algorithms such as logistic regression, decision trees, random forests, support vector machines, and gradient boosting can be effectively used for the prediction and classification of cardiovascular risk (Ambrish et al, 2022) (Ghiasi et al, 2020) (Sumwiza et al, 2023). Furthermore, it is known that processes

performed during the data preprocessing stage, such as normalization, standardization, imputation of missing data, and coding of categorical variables, directly affect model performance (Chew et al, 2025) (Muthumani & Akilandeswari, 2024). Therefore, improving data quality and selecting appropriate data transformation strategies are fundamental requirements for increasing prediction accuracy and model generalizability.

In the analysis of health data, it is not only important to achieve high accuracy rates, but also for the results to be clinically interpretable (García-Vicente et al, 2023). The decision mechanism of a model to be used in clinical practice being transparent and understandable increases the confidence of healthcare professionals in these systems and facilitates their integration into daily practice. In this context, relatively more explainable algorithms such as decision trees or random forests are among the prominent methods in the field of medicine.

In general, it is seen that the development of artificial intelligence and machine learning-based models not only contributes to individual patient management but also significantly supports the creation of policies and strategies aimed at protecting public health (Mooney & Pejaver, 2018) (Morgenstern et al, 2020). Early diagnosis and accurate risk classification enable preventive measures to be implemented more effectively and in a targeted manner, while reducing the workload of healthcare institutions and accelerating clinical decision-making processes. In this context, a detailed review of existing studies in the literature provides a basis for identifying the strengths and limitations of different approaches and for the scientific positioning of this study.

The studies in the literature were comprehensively reviewed, and the findings are summarized below.

Shah et al. (2020) compared Naïve Bayes, decision tree, K-nearest neighbors (KNN), and random forest algorithms for heart disease prediction using the UCI Cleveland dataset (303 samples, 14 variables). In the analysis using basic clinical variables such as blood pressure, chest pain type, and ECG findings, KNN achieved the highest accuracy rate, especially for k=7, demonstrating that heart disease prediction is possible even with a relatively limited number of features (Shah et al, 2020).

Ali et al. (2019) proposed a hybrid model (χ^2–DNN) combining deep neural networks (DNN) with χ^2-based feature selection on the Cleveland dataset. Developed using 297 examples with complete data and 13 basic features, the model achieved 93.33% accuracy and superior results in multiple performance metrics compared to traditional ANN and DNN structures, demonstrating that the combined use of feature selection and deep learning can improve diagnostic accuracy (Ali et al, 2019).

Kavitha and Kaulgud (2023) compared classical K-means with quantum K-means algorithms using the Kaggle Heart Disease UCI dataset (1,025 patients, 13 features). In the study, which used age, gender, chest pain type, cholesterol, and ST segment characteristics, quantum K-means outperformed the classical approach with lower time complexity and better accuracy/evaluation times, thus highlighting the potential of quantum machine learning in medical data mining (Kavitha & Kaulgud, 2023).

Upadhyay et al. (2023) compared the supervised machine learning algorithms KNN, logistic regression (LR), and support vector machines (SVM) on the Cleveland dataset and evaluated their performance using the ROC-AUC metric. The findings revealed that the best

performance was achieved with the LR model, with an AUC value of 0.87, and that the characteristics of the dataset were decisive in model selection (Upadhyay et al, 2023).

Nadeem et al. (2021) developed a hybrid decision-level fusion model that combines an SVM-based classifier with fuzzy logic on two different Kaggle-based datasets (1,025 examples/13 features and 70,000 examples/11 features). The proposed SVM–fuzzy logic architecture achieved an accuracy rate of 96.23%, and testing on different datasets demonstrated the model's generalizability and the role of fuzzy fusion in improving diagnostic accuracy (Nadeem et al, 2021).

Haq et al. (2020) proposed a machine learning-based system for diabetes diagnosis in an e-health environment using the Kaggle Diabetes dataset. The model, which combines an ID3-based feature selection with a decision tree classifier, showed that features such as plasma glucose, DPF, and BMI are decisive, achieved accuracy values above 99% in hold-out, K-fold, and LOSO validations, and provided a high-performance decision support system (Haq et al, 2020).

Singh and Kumar (2020) divided the UCI Heart Disease dataset into 73% training and 27% testing to compare the accuracy of KNN, decision tree, linear regression, and SVM algorithms. In the analyses, KNN produced the best results with 87% accuracy and emerged as a simple yet effective method for predicting heart disease (Singh & Kumar, 2020).

Bharti et al. (2021) compared machine learning and deep learning algorithms on the UCI heart disease dataset, which consists of Cleveland, Hungary, Switzerland, and Long Beach V sub-databases. In the analyses performed with 14 basic features after outlier cleaning (Isolation Forest) and normalization, the artificial neural network-based deep learning model achieved 94.2% accuracy) (the model performance was verified in detail with statistical tests and metrics such as ROC-AUC (Bharti et al, 2021).

Chang et al. (2022) developed a Python-based heart disease detection system on the UCI Heart Disease dataset. After categorical variables were digitized and feature selection was performed, different algorithms were tested. The Random Forest classifier produced the best result with approximately 83% accuracy and was recommended as a feasible decision support tool for clinicians (Chang et al, 2022).

Spencer et al. (2020) combined the Cleveland, Long Beach VA, Hungarian, and Switzerland datasets available at UCI to create a dataset consisting of 720 examples and 14 features) (After missing value and type conversion, they evaluated different feature selection methods (PCA, Chi-Squared, ReliefF, Symmetrical Uncertainty) and various classifiers together. The Chi-Squared + BayesNet combination yielded the best result with 85% accuracy, while different combinations stood out in terms of recall) (chest pain type (cp) was identified as the most predictive attribute (Spencer et al, 2020).

Ahsan et al. (2021) jointly examined 11 machine learning algorithms and 6 different data scaling methods on the UCI Heart Disease dataset, emphasizing the importance of preprocessing steps in clinical data with incomplete and heterogeneous structures. When the CART algorithm was used with Robust Scaler or Quantile Transformer, accuracy and F1 scores of up to 100% were achieved) (it was shown that a single scaling method is not optimal for all algorithms, but that appropriate combinations can dramatically improve performance (Ahsan et al, 2021).

Priyadarshinee and Panda (2022) compared RF, ET, NB, KNN, J48, DTNB, Optimized Forest, and ADTree algorithms after addressing class imbalance with the SMOTE method on the Kaggle heart failure dataset (299 patients, 13 attributes). In experiments conducted in the WEKA environment, the DTNB + SMOTE combination achieved the highest accuracy at 87.08%) (it was shown that SMOTE and hybrid models significantly improved survival prediction performance in imbalanced clinical data (Priyadarshinee & Panda, 2022).

This literature review shows that machine learning methods in the field of heart disease are largely concentrated around supervised classification models and focus on improving performance metrics such as accuracy, sensitivity, and specificity. A significant portion of the studies compares algorithms such as logistic regression, K-nearest neighbors, support vector machines, random forests, and deep neural networks) (a more limited portion is devoted to the effects of feature selection, class imbalance, and data scaling strategies. However, the identification of latent patient subgroups in clinical heart disease data, the systematic comparison of different clustering methods, and the detailed examination of the effects of distance measures and data transformation/scaling approaches on cluster structure have remained relatively limited.

This study aims to fill this gap in the literature by focusing on unsupervised learning approaches on clinical data related to heart disease, comparatively evaluating different clustering algorithms, distance measures, and data scaling methods. The next section presents the characteristics of the dataset used in the study, the preprocessing steps, and the applied methodological framework in detail, thus laying the groundwork for the analysis of the findings.

Methodology

This section details the dataset used in conducting the study, the preprocessing steps applied, and the methodological approach adopted. First, the structure, variables, and basic statistical properties of the dataset used in the analyses are introduced) (then, the preprocessing steps performed to make the data suitable for analysis are explained. Subsequently, the algorithms preferred in cluster analysis, the distance measures and data scaling approaches used, and the evaluation criteria considered in comparing clusters are presented systematically. This aims to present the methodological framework followed in the study in a clear, consistent, and reproducible manner.

I. Data Set and Detailed Analysis of Data

The analysis utilized a heart disease dataset from patients monitored at the VA Medical Center in Long Beach, California. The dataset contains the clinical records of a total of 303 patients, with each individual represented by 11 clinical variables. These variables include age, gender, type of chest pain, resting blood pressure, serum cholesterol level, fasting blood sugar, electrocardiogram (ECG) findings, maximum heart rate, presence of exercise-induced angina, ST segment depression (oldpeak), and ST segment slope, which are basic cardiovascular parameters. The dataset includes 206 men and 97 women) (the patients' ages range from 29 to 77 years. The general structure and basic descriptive statistics of the dataset are summarized in Table 1.

Table 1. Definition and value ranges of clinical variables used in the dataset

Variable	Definition	Number of Samples	Minimum Value	Maximum Value
age	Patient's Age	303	29	77
sex	Patient's Gender	303	0	1
cp	Type of Chest Pain	303	1	4
trestbps	Resting blood pressure (in mm Hg upon hospital admission)	303	94	200
chol	Serum cholesterol in mg/dl	303	126	564
fbs	Fasting blood sugar > 120 mg/dl (1=true) (0=false)	303	0	1
restecg	Resting electrocardiogram results	303	0	2
thalach	Maximum heart rate achieved	303	71	202
exang	Exercise-induced angina (1 = yes) (0 = no)	303	0	1
oldpeak	Numerical value measured during depression	303	0	6,2
slope	Slope	303	1	3

The dataset used consists of clinical characteristics associated with heart disease. A systematic preprocessing stage was applied to the data before proceeding to the analysis process. In this context, categorical and binary (boolean) variables were excluded from the analysis because the k-means clustering algorithm produces more consistent results on continuous and numerical variables. Including these variables directly in the k-means algorithm has the potential to weaken the meaningfulness of distance calculations and the interpretability of distinctions between clusters. During the preprocessing stage, observations with missing values were also excluded, and only records containing complete data were included in the analysis. This enhanced the reliability of the calculated distances and ensured that the cluster structure was examined on a more statistically robust basis. The resulting dataset, obtained through these steps, was transformed into a homogeneous and numerically weighted structure consistent with the assumptions of the k-means algorithm. This strengthened the validity and clinical interpretability of the findings from the clustering analysis. Figure 1 shows histograms and density curves for the variables of age, resting blood pressure (trestbps), serum cholesterol (chol), maximum heart rate (thalach), and ST depression (oldpeak), along with summary statistics (mean, standard deviation, quartiles, kurtosis, skewness).

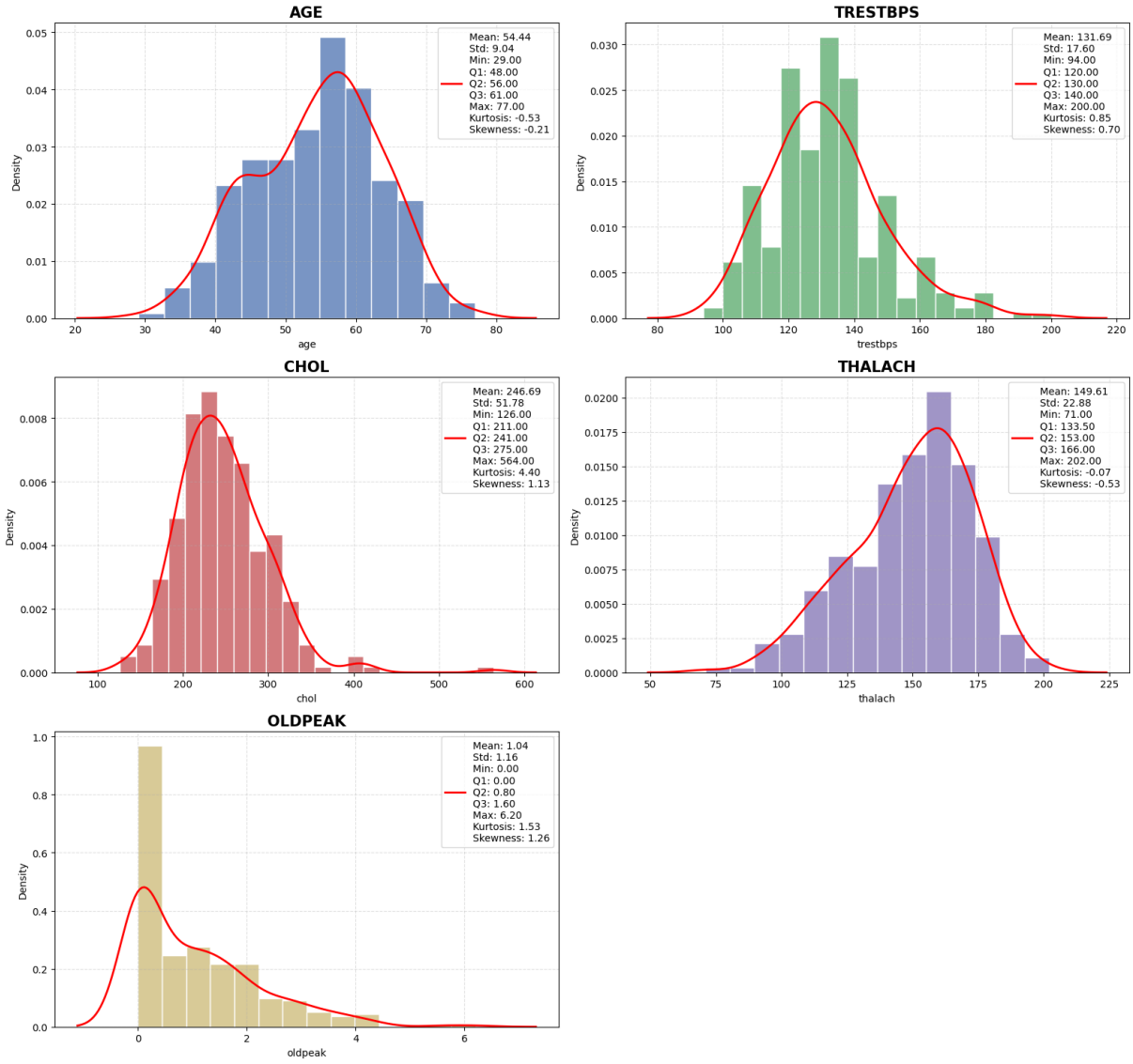

Figure 1. Distribution and Summary Statistics of Key Clinical Variables in the Heart Disease Dataset

Figure 1 visually represents the statistical distributions of the primary clinical variables used in the study. The age distribution indicates that the sample predominantly consists of middle-aged and elderly individuals, while the trestbps (DeGuire et al, 2019) and chol variables generally exceed the limits considered normal and, due to positive skewness, reach quite high values in some cases. The Thalach variable is concentrated more in the middle-high range, while the oldpeak variable shows low ST depression in most patients and high ST depression in a small number of patients. These findings indicate that the distributions in the dataset are not homogeneous and that outliers are present) (therefore, scaling and standardizing the variables is necessary before proceeding to k-means clustering analysis. Figure 2 shows the bivariate correlation coefficients between the variables of gender (sex), fasting blood sugar (fbs), and exercise-induced angina (exang).

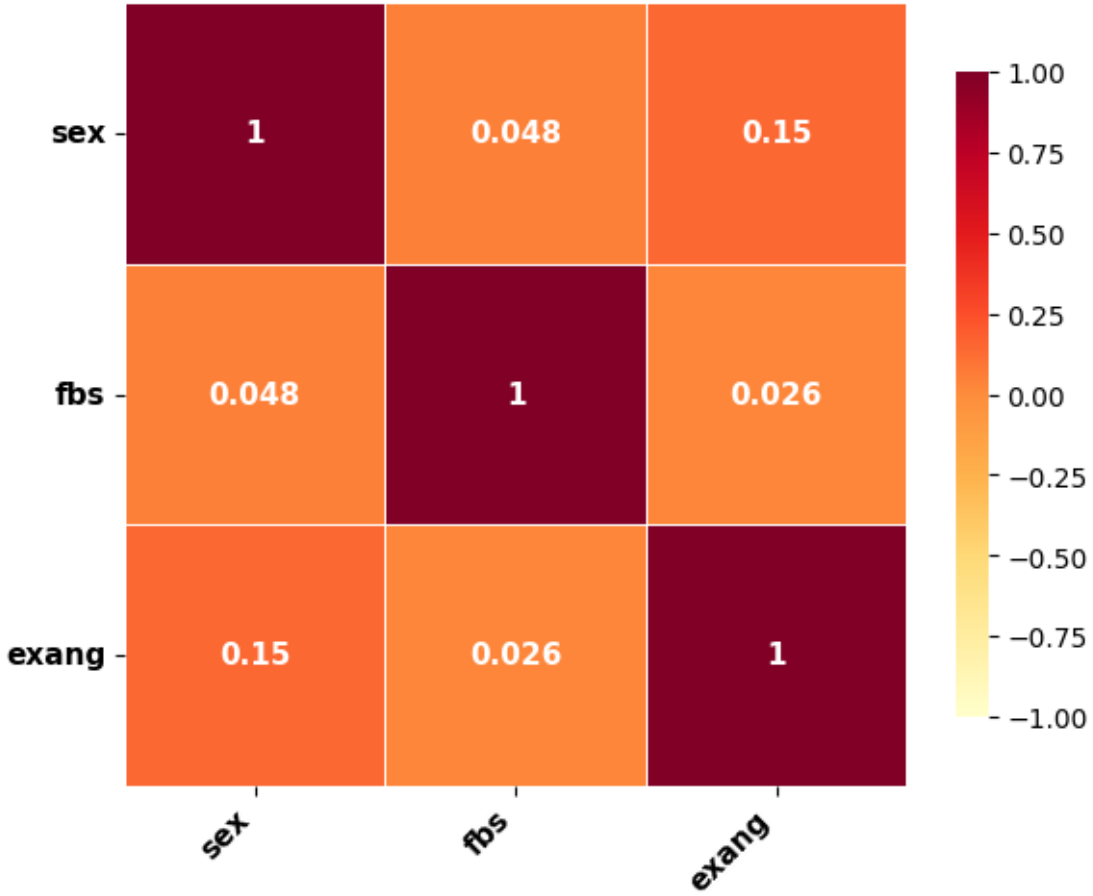

Figure 2. Heat Map of Correlation Coefficients Between Binary Clinical Indicators.

According to Figure 2, the fact that the absolute value of all coefficients is below 0.15 indicates that there are only weak linear relationships between the variables. Although the correlation between gender and exercise-induced angina is relatively higher than other pairs, it does not reach a significant level of dependence. These findings reveal that the aforementioned dual clinical indicators are largely independent of each other and do not pose a significant risk in terms of multiple linear dependence. Figure 3 shows the Pearson correlation coefficients between the numerical variables used in the study.

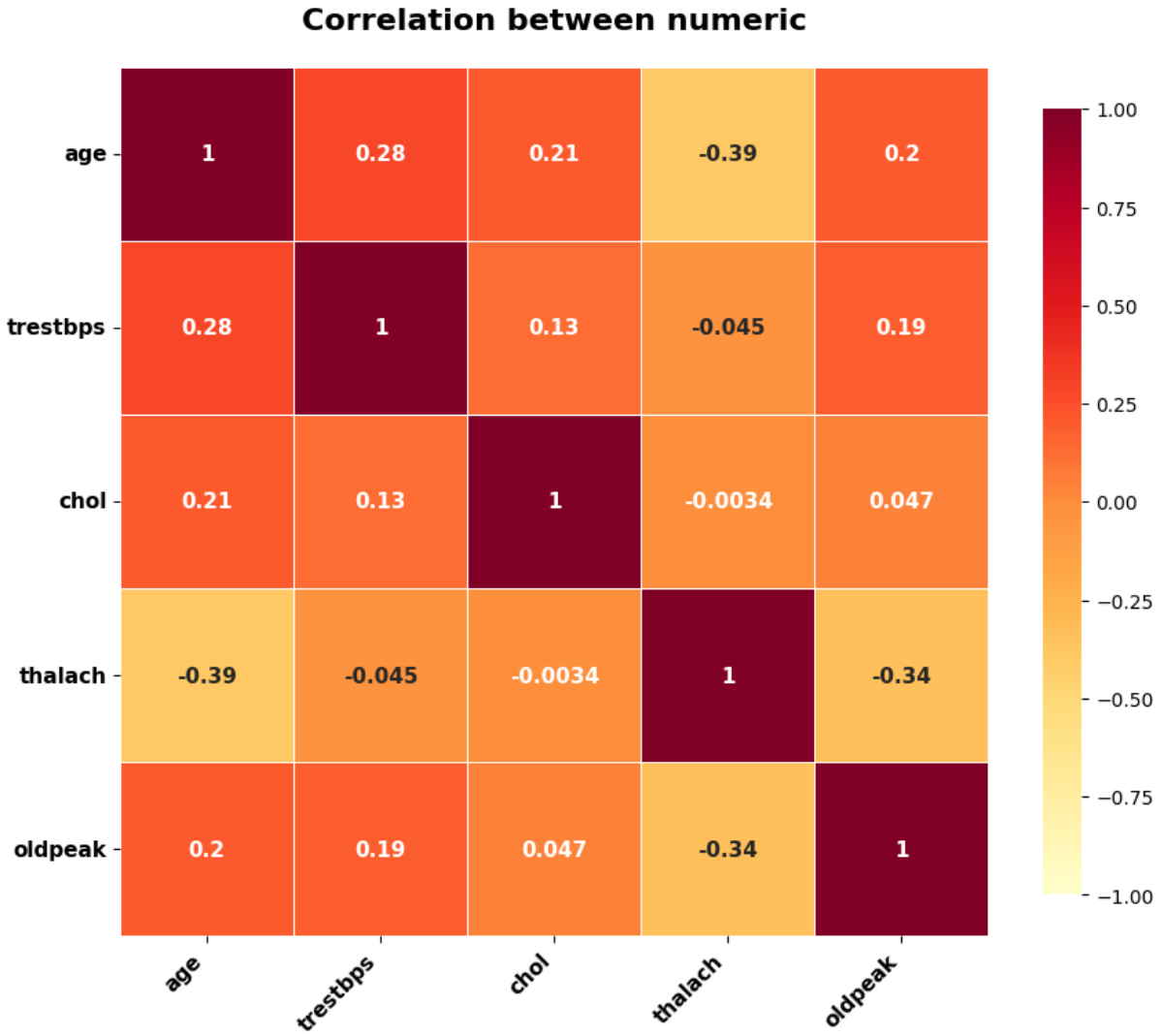

Figure 3. Pearson correlation matrix between clinical variables.

According to Figure 3, the results show a moderate negative correlation (r = -0.39) between age and maximum heart rate (thalach), indicating that the maximum heart rate achievable decreases with increasing age. The negative relationship between thalach and ST depression (oldpeak) (r = -0.34) is similarly noteworthy. Apart from age showing a weak positive correlation with resting blood pressure (trestbps) and cholesterol (chol), the other coefficients are low, and there are no strong linear relationships between the variables. This suggests that the variables can contribute relatively independently in the modeling process and that the risk of multicollinearity is limited. Equation (1) presents the mathematical expression used to calculate the Pearson correlation coefficient (Zhou et al, 2016).

$$r_{xy} = \frac{\sum(x_i - \bar{x})\sum(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \quad (1)$$

The symbols used in Equation (1) are defined as follows:

- x_i and y_i: are the i. observation values of the variables x and y, respectively.
- $\bar{x} = \frac{1}{n}\Sigma_{i=1}^{N} x_i$: is the arithmetic mean of the variable x.

- $\bar{y} = \frac{1}{n} \Sigma_{i=1}^{N} y_i$: the arithmetic mean of the y variable.
- The numerator represents the covariance between the two variables.
- The denominator is the product of the standard deviations of both variables.

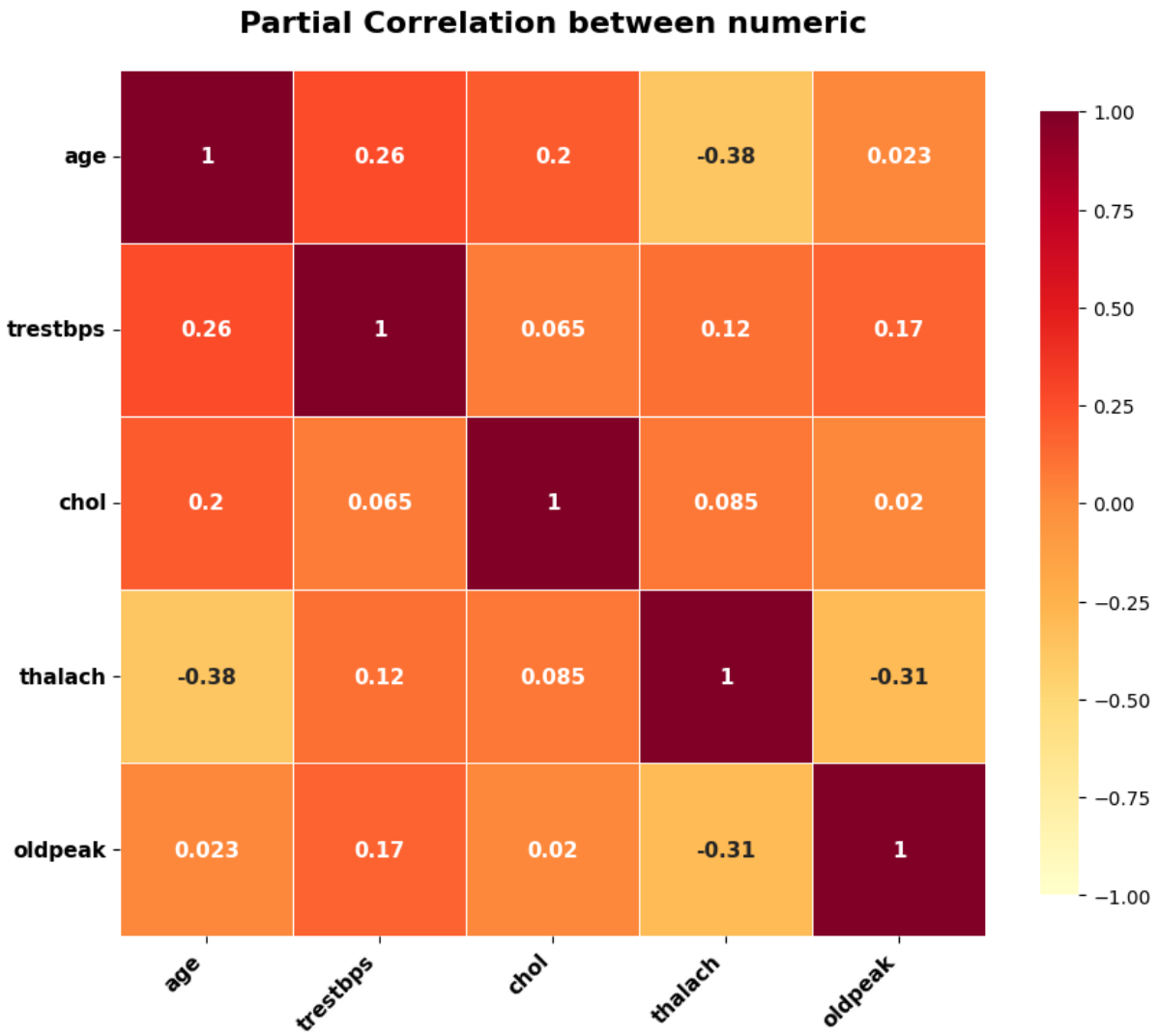

Figure 4. Partial correlation matrix between clinical variables.

According to Figure 4, partial correlation measures the strength of the specific relationship between two variables while controlling for the effects of other variables. The fact that the results are lower than Pearson correlations indicates that some relationships are partially explained by other variables. In particular, the negative relationship between age and maximum heart rate (thalach) (r = -0.38) and the negative relationship between thalach and ST depression (oldpeak) (r = -0.31) are clinically significant. The low coefficients observed in other variable pairs suggest that there is no significant multiple linear regression problem and that the independent variables are at an acceptable level for statistical analysis and modeling.

Equation (2) provides the mathematical definition of the partial correlation between x_i and x_j when the variable x_k is held constant (Kim, 2015).

$$r_{ij|k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{1 - r_{ik}^2}\sqrt{1 - r_{jk}^2}} \tag{2}$$

Equation (3) provides the mathematical definition of the semi-partial correlation between x_i and x_j when the variable x_k is controlled (Kim, 2015).

$$r_{i(j|k)} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{1 - r_{jk}^2}} \tag{3}$$

Figure 5 shows the distribution of numerical variables (age, trestbps, chol, thalach, oldpeak) and possible outliers using box plots.

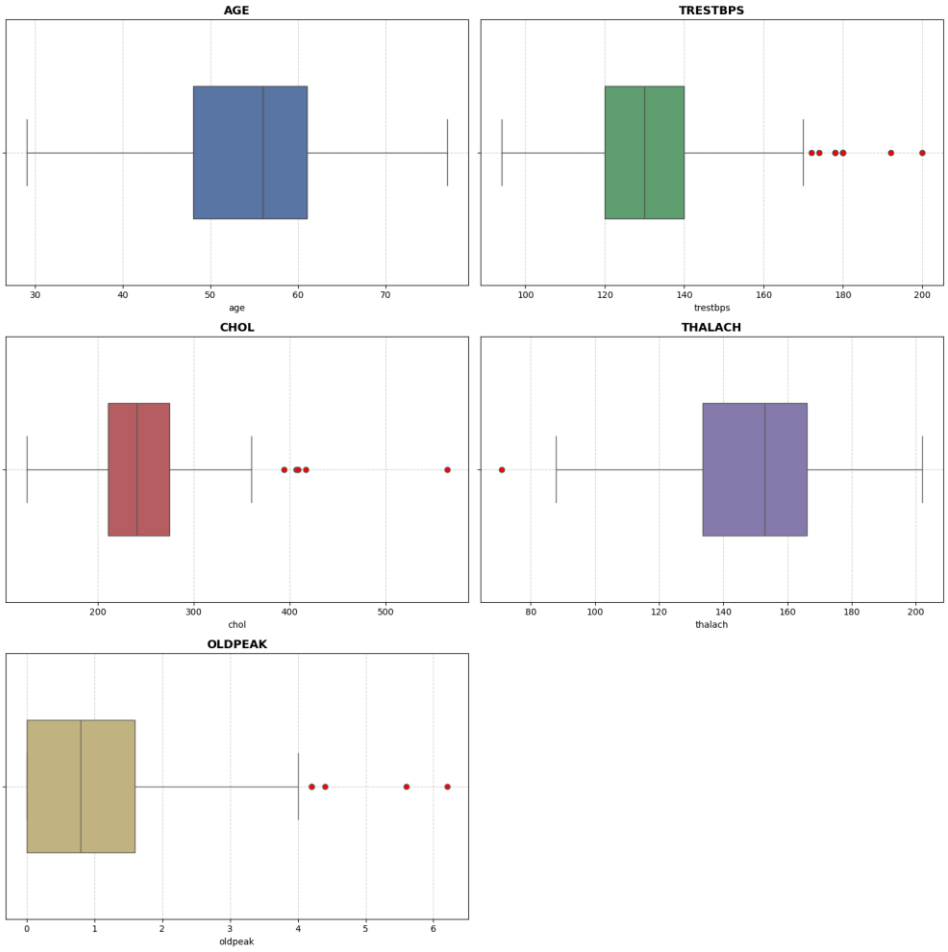

Figure 5. Box plots of clinical variables and visual inspection of potential outliers

According to Figure 5, the median and Interquartile Range (IQR) are summarized in box plots, and observations outside the IQR limits are marked as outliers. The presence of multiple extreme values in particular variables such as trestbps, chol, and oldpeak indicates

that the variance in these variables is relatively high and that outliers should be considered before modeling.

In machine learning and statistical modeling, variables defined on different scales can negatively affect the performance of distance-based (k-means, KNN, etc.) and gradient-based methods. To prevent large-scale variables from overshadowing others, scaling was applied during the data preprocessing stage, and two common scaling methods were preferred in this study.

Normalization aims to transform each variable into a specific range. In practice, the [0, 1] range is often preferred, and the transformation used is defined in Equation (4) (Patro & Sahu, 2015).

$$A' = \left(\frac{A - min\ value\ of\ A}{max\ value\ of\ A - min\ value\ of\ A}\right) * (D - C) + C \quad (4)$$

In Equation (4):

- A' represents the data after Min-Max normalization has been applied.
- If predefined limits are given in the range [C,D], the data is rescaled to this range.
- A indicates the value range in which the original data is defined.

This method ensures that normalized values are obtained by converting the values in the given original range to the [C,D] range. With this method, the data is rescaled so that the smallest value is 0 and the largest value is 1. This makes variables measured in different units comparable.

Like min-max normalization, z-scores (standardization) also provide a value range between 0 and 1. This transformation is calculated using the expression shown in Equation (5) (Patro & Sahu, 2015).

$$v_i' = \frac{v_i - \bar{E}}{std(E)} \quad (5)$$

Equation (5):

- v_i': Standardized (z-score) value
- v_i: Original value for the i. observation
- $\bar{E}$: Arithmetic mean of the corresponding column (feature)
- $std(E)$: Standard deviation of the corresponding column

The standard deviation is defined as shown in Equation (6) (Patro & Sahu, 2015).

$$std(E) = \sqrt{\frac{1}{(n-1)}\sum_{i=1}^{n}(v_i - \bar{E})^2} \quad (6)$$

This method produces more appropriate and consistent results, especially for data with a distribution close to normal) (it also contributes to more stable and reliable parameter estimation in regression models (Ali, 2022). Both standardization and normalization were applied to all numerical variables in the dataset. The resulting transformed values were used to ensure more stable operation of the k-means clustering algorithm in the subsequent

stage. This eliminated scale differences between variables and aimed to balance the relative contribution of each variable during the model's learning process.

II. Hierarchical Clustering Methods

At this stage, hierarchical clustering analyses were performed to evaluate which distance metric and which clustering approach would be more appropriate before proceeding to the k-means algorithm. Since not only Euclidean distance but also cosine distance is among the commonly used metrics in medical data analysis, both distance metrics were considered comparatively (Ni et al, 2018) (Prasetyo et al, 2023). The cosine distance is a distance measure based on the angle between vectors in a multidimensional space, emphasizing the directional similarity of variables rather than magnitude. Figure 6 schematically illustrates the different linkage criteria used in hierarchical clustering.

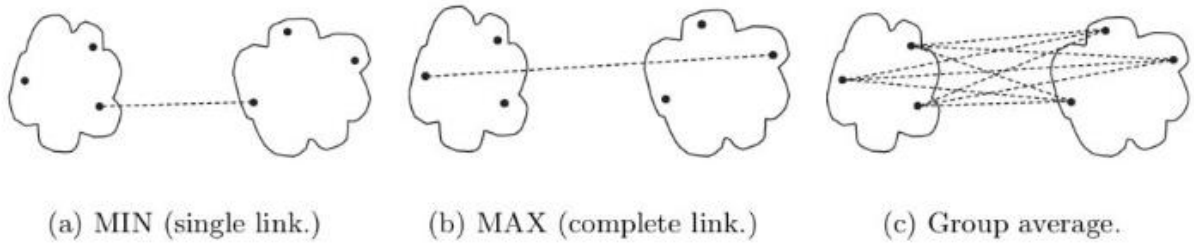

Figure 6. Schematic representation of single bond, full bond, and group average bond criteria in clustering (Jarman, 2020).

According to Figure 6, in the single linkage approach, clusters are combined based on the distance between the two closest points) (in the complete linkage approach, they are combined based on the distance between the two farthest points. In the group average method, the average distance calculated for all point pairs between clusters is used as the basis. The dendrograms in Figure 7 show how different linkage criteria used in hierarchical clustering affect the cluster structure.

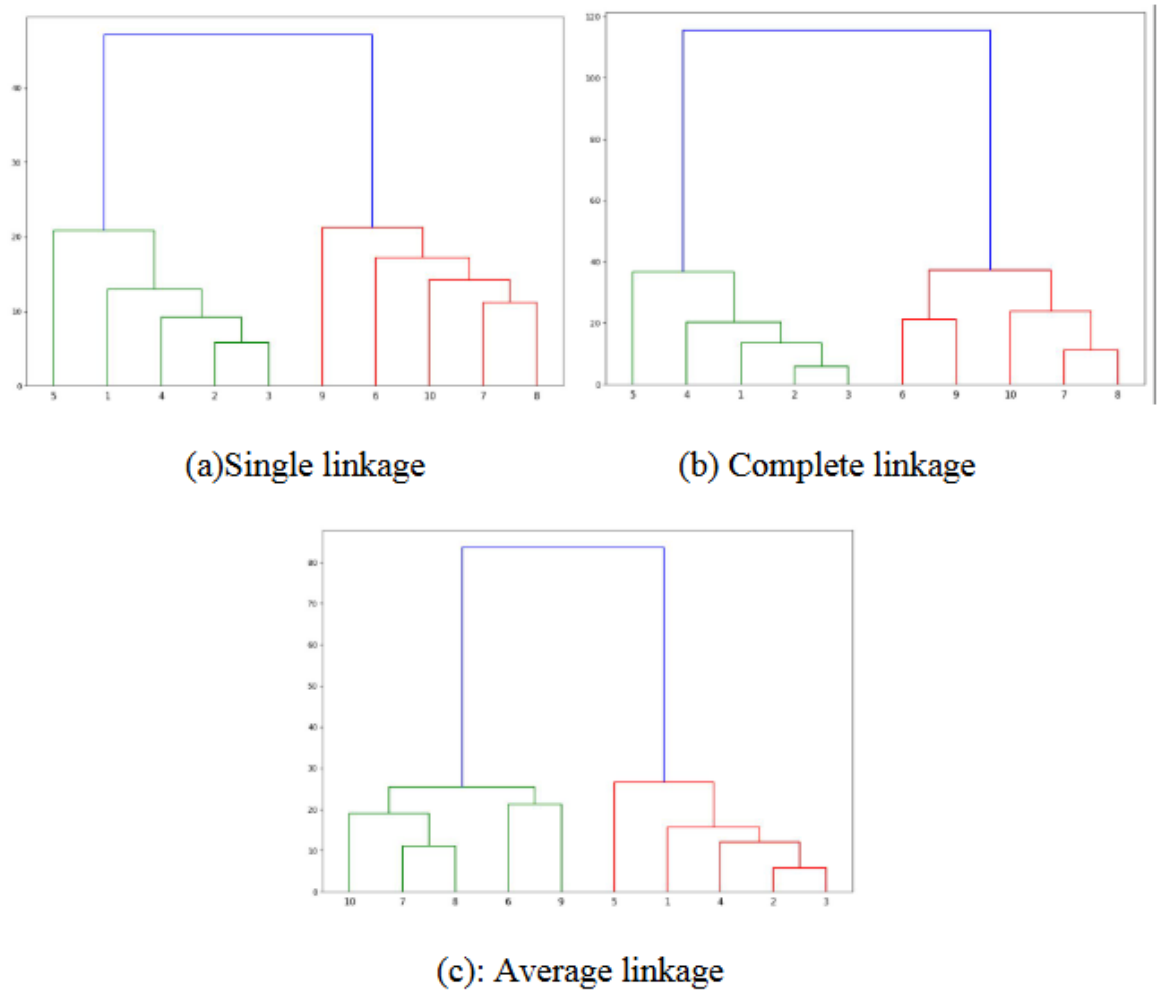

Figure 7. Comparative view of dendrograms obtained from single linkage, complete linkage, and average linkage hierarchical clustering approaches (Jarman, 2020).

According to Figure 7, single linkage is more prone to chaining effects because it is based on the shortest distance between clusters. Complete linkage produces more compact clusters by using the farthest distance, while average linkage offers a more balanced structure between these two extremes by using the average distance between all point pairs.

a. Single Linkage Method

The single linkage method is a hierarchical clustering approach that operates from the bottom up. The distance between two clusters is defined as the smallest distance between any element in one cluster and any element in the other cluster. Initially, each observation is considered a separate cluster) (in each iteration, the two clusters with the smallest distance between them are merged. Then, the smallest distances between the newly formed cluster and the other clusters are recalculated, and the cluster pair with the shortest distance is merged again. This process continues until all observations are grouped under a single hierarchical structure (Sharma & Batra, 2019). The dendrograms presented in Figures 8-11 show that the Single Linkage (nearest neighbor) method does not produce a distinct and

interpretable cluster structure in either standardized or normalized data, using either Euclidean or cosine distance metrics.

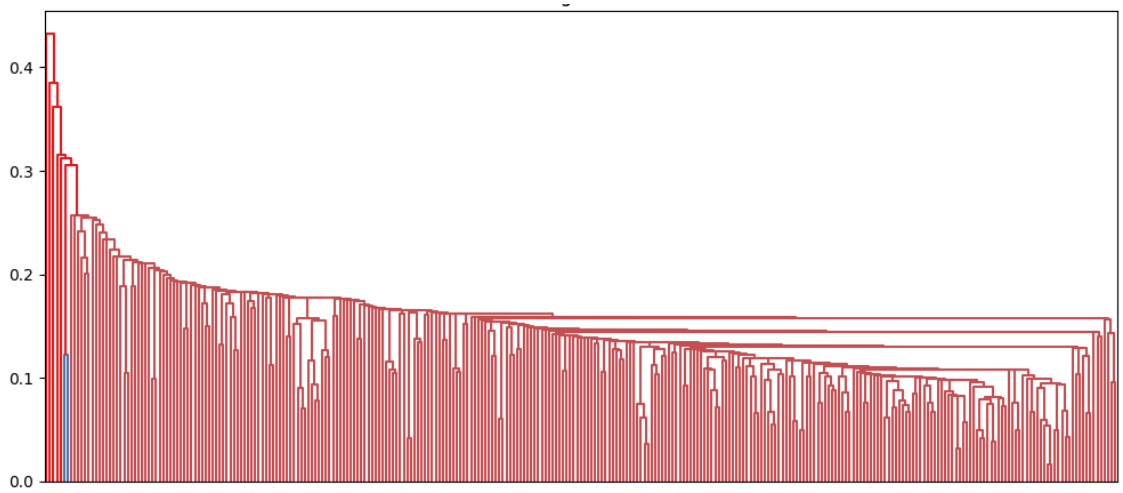

Figure 8. Hierarchical clustering dendrogram obtained using Euclidean distance and single linkage method on normalized data

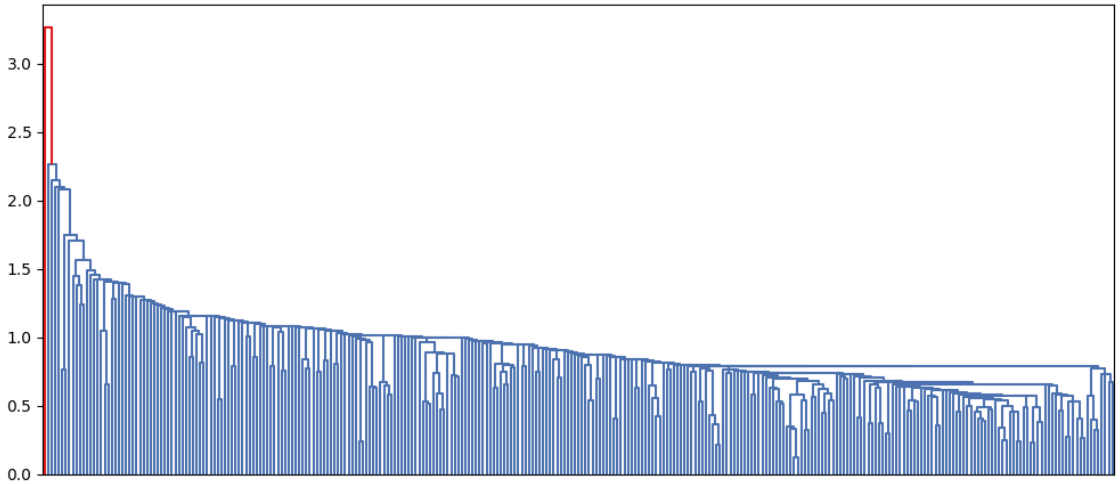

Figure 9. Hierarchical clustering dendrogram obtained using Euclidean distance and single linkage method on standardized data

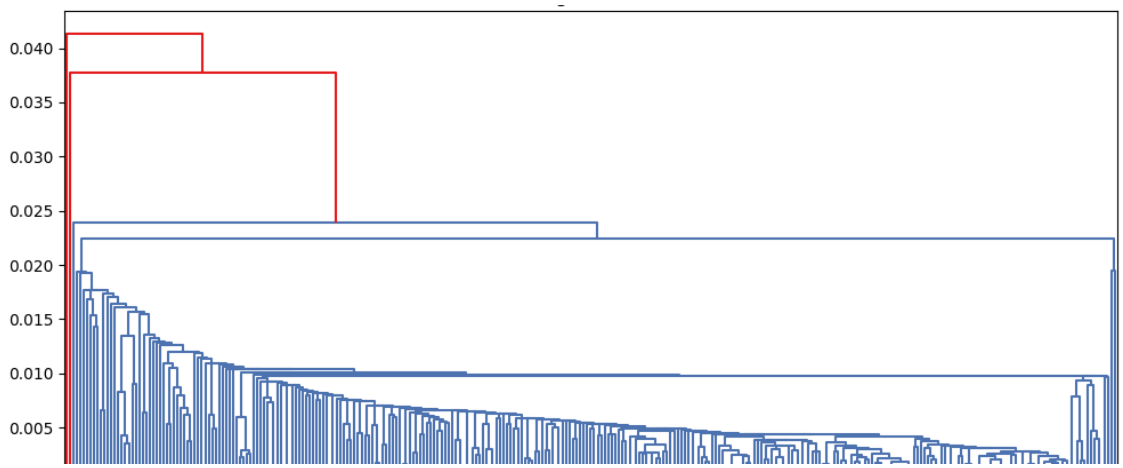

Figure 10. Dendrogram obtained using the single linkage method on normalized data based on cosine distance

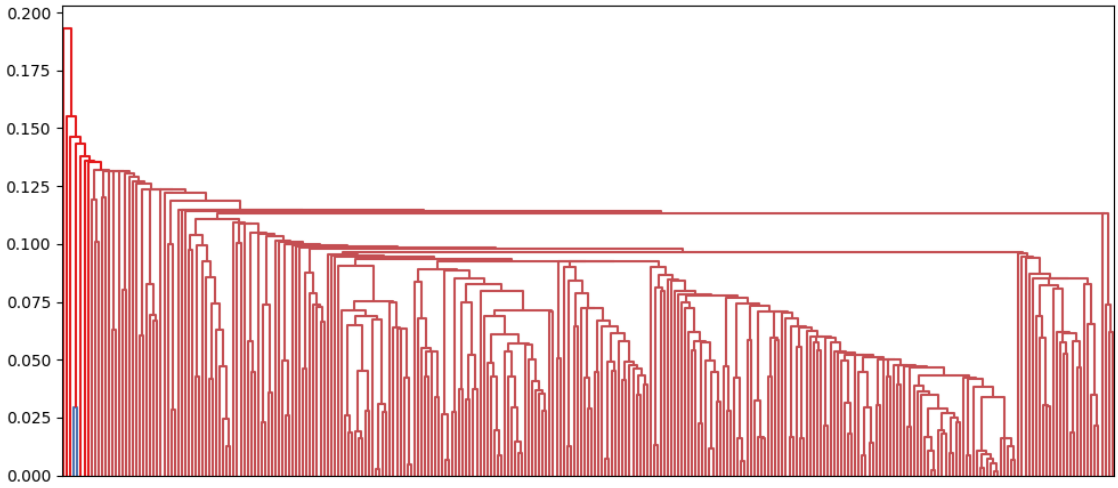

Figure 11. Hierarchical clustering dendrogram obtained using cosine distance and single linkage method on standardized data

As emphasized in the literature, this situation stems from the method's tendency to chain, combining small clusters sequentially to form long chains (Jarman, 2020). This tendency causes the actual separation between clusters to become blurred and distorts the natural structure of the data, particularly in high-dimensional medical data. Therefore, in standardized and normalized datasets, the Single Linkage approach is not considered a suitable method for obtaining reliable and meaningful subgroup structures in clinically critical applications such as heart diseases.

b. Average Linkage Method

Average linkage is also referred to in the literature as the Unweighted Pair Group Method with Arithmetic Mean (UPGMA). This approach, proposed by Sokal and Michener (Sokal & Michener, 1958) to overcome the limitations of single and complete linkage methods, defines the distance between two clusters as the arithmetic mean of the distances between each element in one cluster and each element in the other cluster. Thus, the distance between clusters is evaluated in a more balanced manner, providing a natural compromise between linkage criteria (Yim & Ramdeen, 2015). The dendrograms in Figures 12-15 show that the average linkage method creates more distinct and balanced cluster structures on the data compared to single linkage.

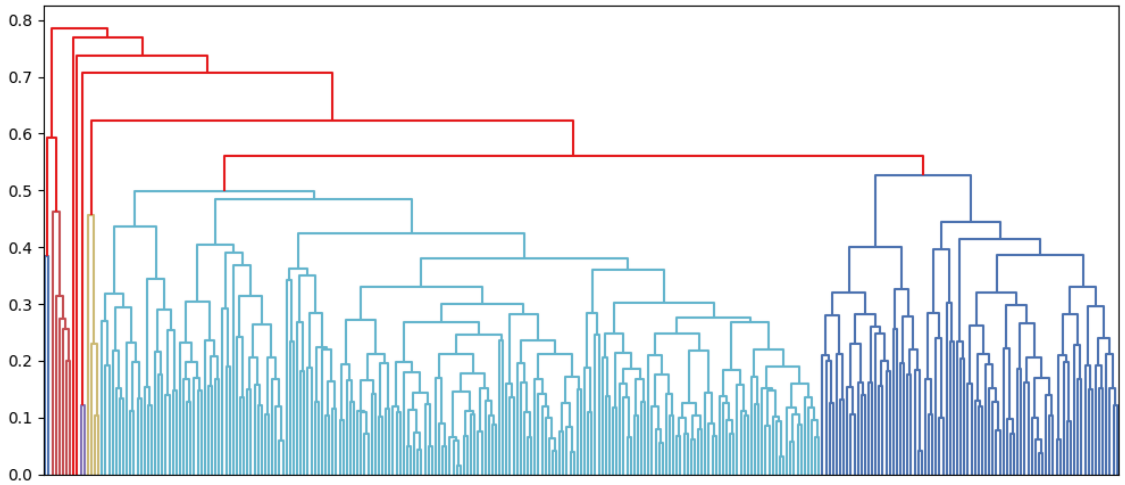

Figure 12. Hierarchical clustering dendrogram obtained using Euclidean distance and average linkage method on normalized data

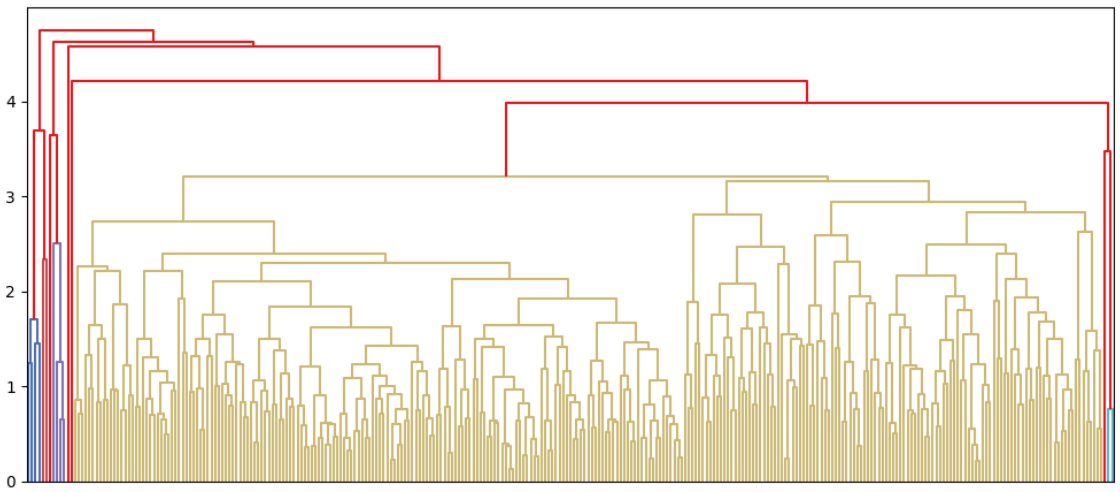

Figure 13. Hierarchical clustering dendrogram obtained using Euclidean distance and average linkage method on standardized data

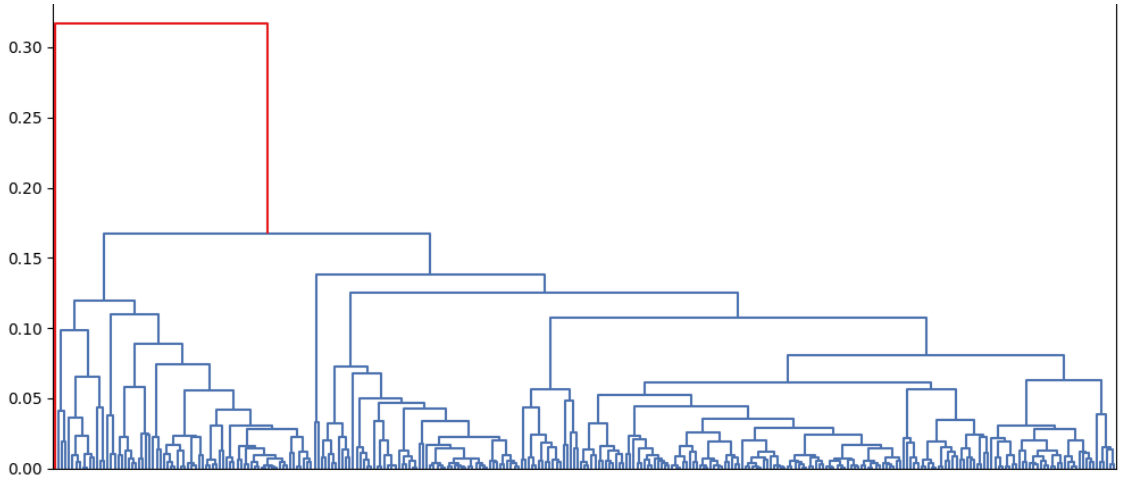

Figure 14. Hierarchical clustering dendrogram obtained using cosine distance and average linkage method on normalized data

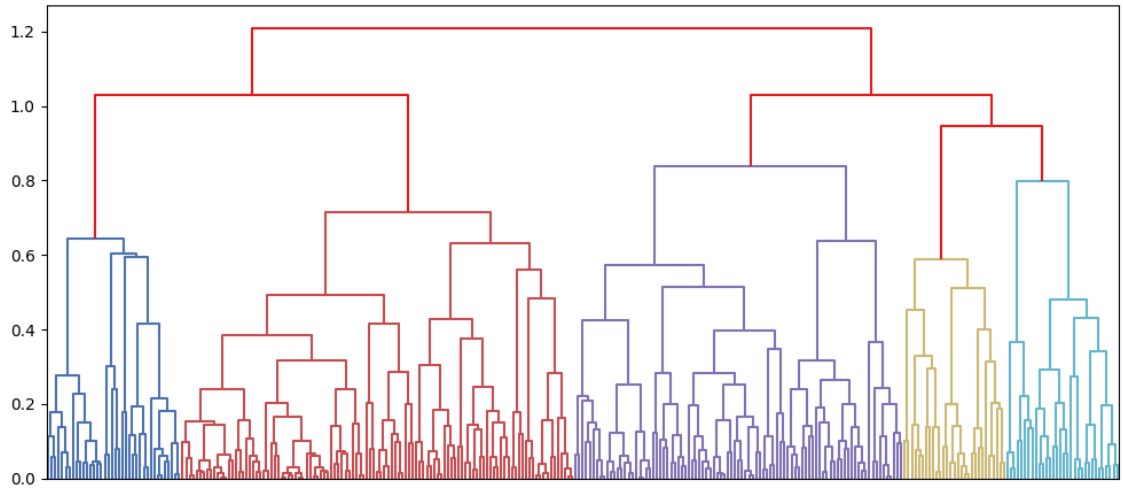

Figure 15. Hierarchical clustering dendrogram obtained using cosine distance and average linkage method on standardized data

According to Figures 12-15, dendrograms show that the average linkage method creates more distinct and balanced cluster structures on the data compared to single linkage. This effect is particularly pronounced when cosine distance and standardized data are used. While Euclidean distance better reflects the absolute value differences between observations, cosine distance provides an advantage in grouping patterns that are different in magnitude but exhibit similar orientations. These findings reveal that the single linkage approach has limitations in health data, while the average linkage method provides more reliable clustering structures, forming the basis for selecting the appropriate distance measure in subsequent k-means analysis.

c. Complete Linkage Method

The complete linkage method is also known as the farthest neighbor or maximum distance method. This approach is similar to the single linkage (nearest neighbor) method, but it evaluates the similarity between clusters based on the farthest distance rather than the nearest distance. While this eliminates the chaining problem, it introduces a different disadvantage. For example, observations a, b, c, and d, which are close to each other according to a specific set of variables, may belong to the same cluster, but the presence of observation e, which has values significantly different from this group, may prevent the clusters from merging. In this case, the furthest distance criterion between clusters becomes decisive, and the effect of outliers is exaggerated, preventing the merging of nearby clusters (Yim & Ramdeen, 2015).

At this stage, hierarchical clustering analysis was applied to the data set. Two distance measures were used in the analysis: Euclidean and cosine distances. The complete linkage method, which is based on the most distant observation pair between clusters, was preferred as the linkage measure. The data was first normalized and then standardized in two different preprocessing scenarios) (in both cases, dendrograms were obtained using the same parameters and visually compared. Thus, the effects of differences in distance measure and preprocessing strategy on the clustering structure were systematically evaluated.

Figure 16 shows the dendrogram obtained using Euclidean distance and the complete linkage method on normalized data.

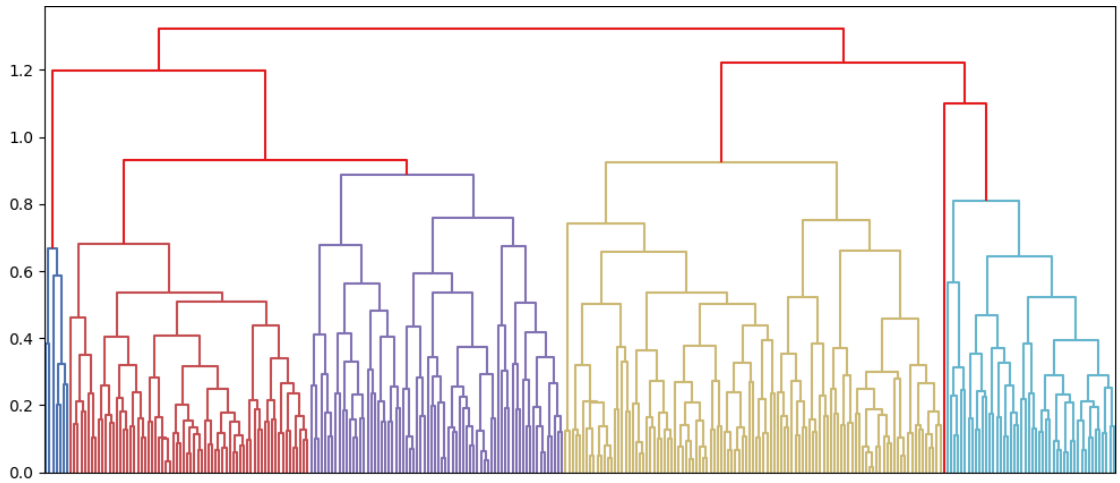

Figure 16. Hierarchical clustering dendrogram obtained using Euclidean distance and complete linkage method on normalized data

When examining the dendrogram in Figure 16, it is evident that the four main clusters are distinctly separated and that the hierarchical structure of the cluster mergers is regular, gradual, and balanced. The early addition of small sub-clusters to higher-level clusters indicates that there are limited outliers in the dataset. These findings show that the four-cluster solution on normalized data provides a stable and interpretable clustering structure. In the dendrogram in Figure 17, it can be seen that the data set is first divided into two main clusters at the top level, and the large cluster on the right is then divided at lower levels, resulting in a total of four dominant clusters.

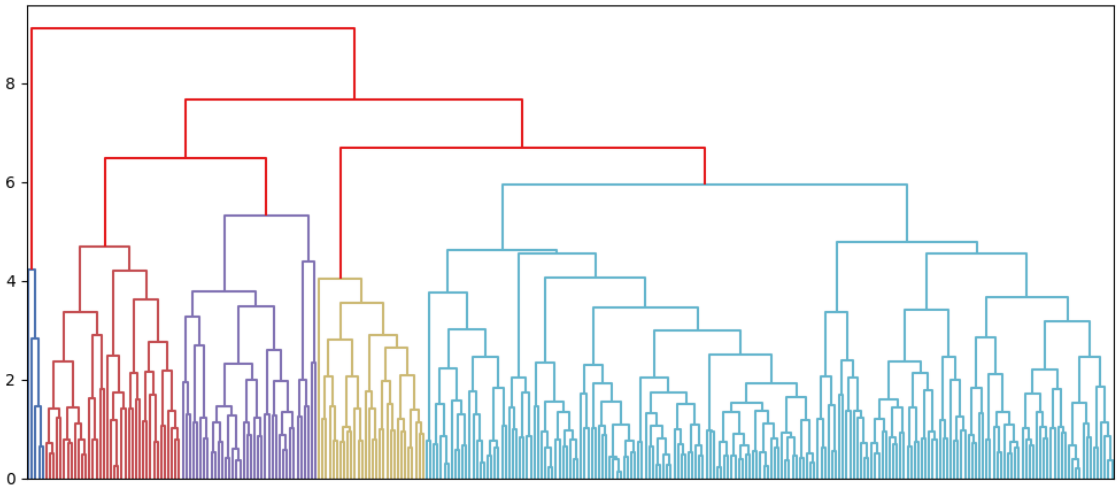

Figure 17. Hierarchical clustering dendrogram obtained using Euclidean distance and complete linkage method on standardized data

Figure 17 suggests that the structure indicates the presence of a hierarchical clustering pattern in the dataset and that a four-cluster solution provides a more explanatory structure upon detailed examination. The dendrogram in Figure 18 shows that

the structure is clearly divided into two main clusters at the top level when cosine distance and complete linkage are used.

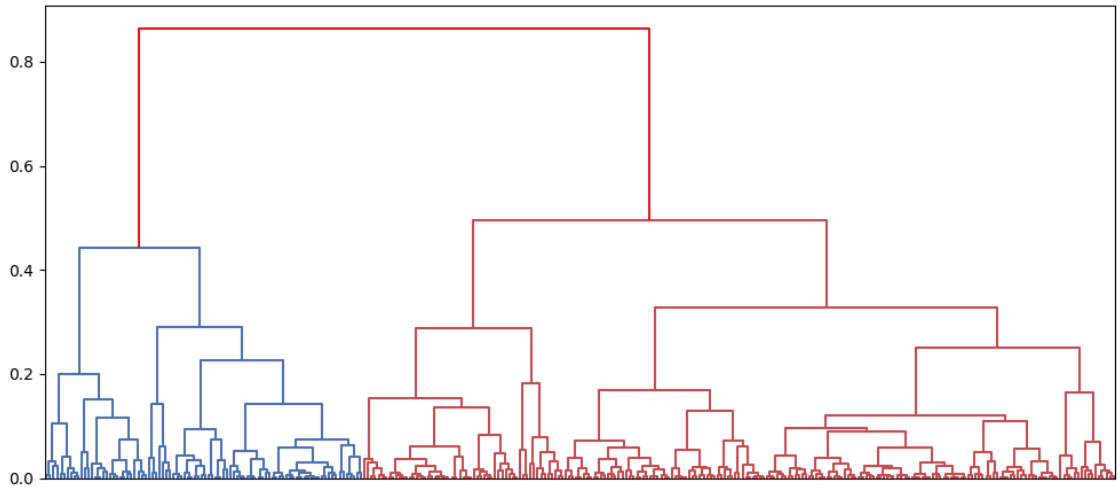

Figure 18. Hierarchical clustering dendrogram obtained using cosine distance and complete linkage method on normalized data

According to Figure 18, dominant binary dissimilarity does not allow for sufficient identification of possible subset structures in the data set) (therefore, the two-cluster solution represents data diversity in an excessively reductive manner from an analytical perspective. The dendrogram in Figure 19 shows that when cosine distance and complete linkage are used on standardized data, a large number of small clusters are formed at the top level, and the cluster structure exhibits a more fragmented appearance compared to previous examples.

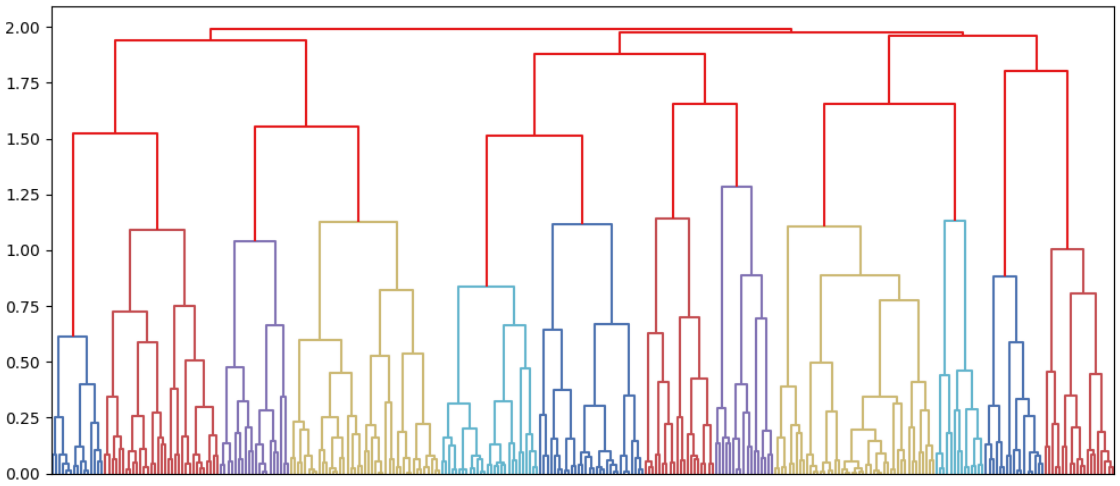

Figure 19. Hierarchical clustering dendrogram obtained using cosine distance and complete linkage method on standardized data

According to Figure 19, this excessive fragmentation, extending to approximately 12 sub-clusters, limits the practical interpretability of the obtained solution by causing observations with similar characteristics to be scattered across different groups and weakening the structural integrity between clusters.

III. Determining the Number of Clusters: Elbow and Silhouette Criteria

Determining the optimal number of clusters in cluster analysis is a critical step for the validity of the results. In this study, two widely used evaluation measures were employed to determine the number of clusters: the Elbow and Silhouette methods. The Elbow method examines the change in the total sum of squared errors (SSE) values calculated for each number of clusters) (it suggests the inflection point where the rate of decrease in the curve slows down significantly as the candidate solution corresponding to the appropriate number of clusters. The Silhouette method, on the other hand, compares the similarity of each observation to its own cluster with its distance to the closest alternative cluster, producing an index ranging from -1 to +1. It quantitatively evaluates the overall separation and cohesion level of the clustering solution based on this value.

a. Elbow Method

Following hierarchical clustering results, the Elbow and Silhouette metrics were used to determine the optimal number of clusters. The Elbow method is based on examining the decrease in the Sum of Squared Errors (SSE) value as the number of clusters k increases. The "elbow" point on the SSE curve plotted against the number of clusters, where the rate of decrease becomes markedly slower, is interpreted as the candidate solution corresponding to the optimal number of clusters. The SSE calculation used in this study is given in Equation (7) (Marutho et al, 2018).

$$SSE = \sum_{K-1}^{K} \sum_{xi \in Sk} \|Xi - Ck\|_2^2 \quad (7)$$

In Equation (7):

- K: denotes the total number of clusters,
- Sk: denotes the observation set belonging to the kth cluster,
- xi: denotes the data points,
- Ck: denotes the centroid of the kth cluster.

SSE represents the sum of the squares of the Euclidean distances of each data point to its own cluster center. As the number of clusters increases, it is natural for SSE to decrease, which stems from the reduction in variance within each cluster. However, increasing the number of clusters beyond a certain value does not yield a significant improvement in SSE. This breakpoint, where the rate of decrease in the SSE–k curve slows down significantly, is called the "elbow" and is used as a key criterion for determining the optimal number of clusters. Figure 20 shows the Elbow graph obtained using cosine distance and complete linkage on normalized data and the determination of the optimal number of clusters.

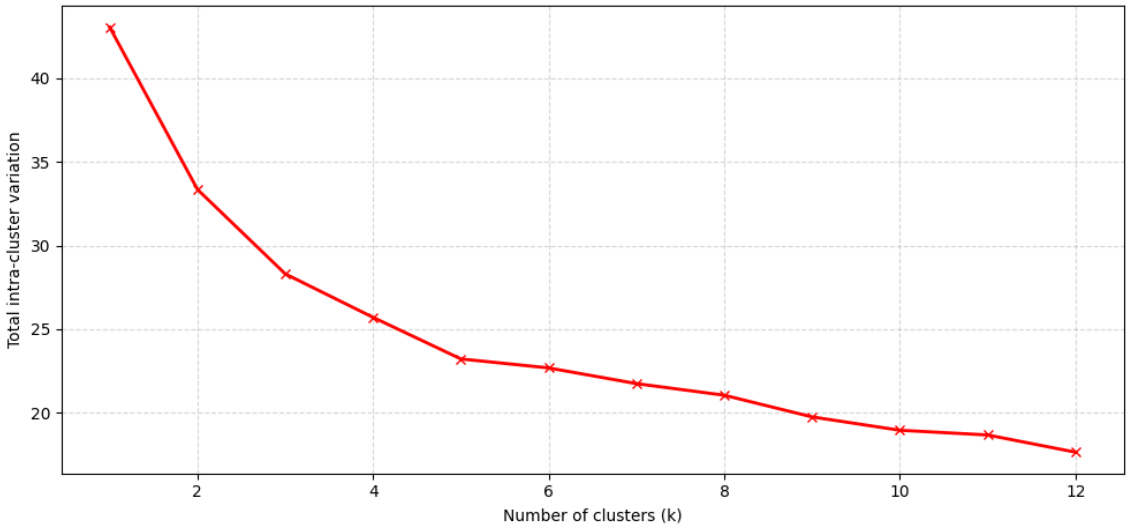

Figure 20. Elbow graph obtained using cosine distance and complete linkage on normalized data and determination of the optimal number of clusters

The Elbow graph in Figure 20 shows that the SSE values decrease significantly, especially up to k=3, and then the curve flattens out. Accordingly, it can be said that a three-cluster solution under the cosine distance metric is an appropriate option for the normalized data. Figure 21 presents the Elbow graph obtained using cosine distance and complete linkage on standardized data, along with the determination of the optimal number of clusters.

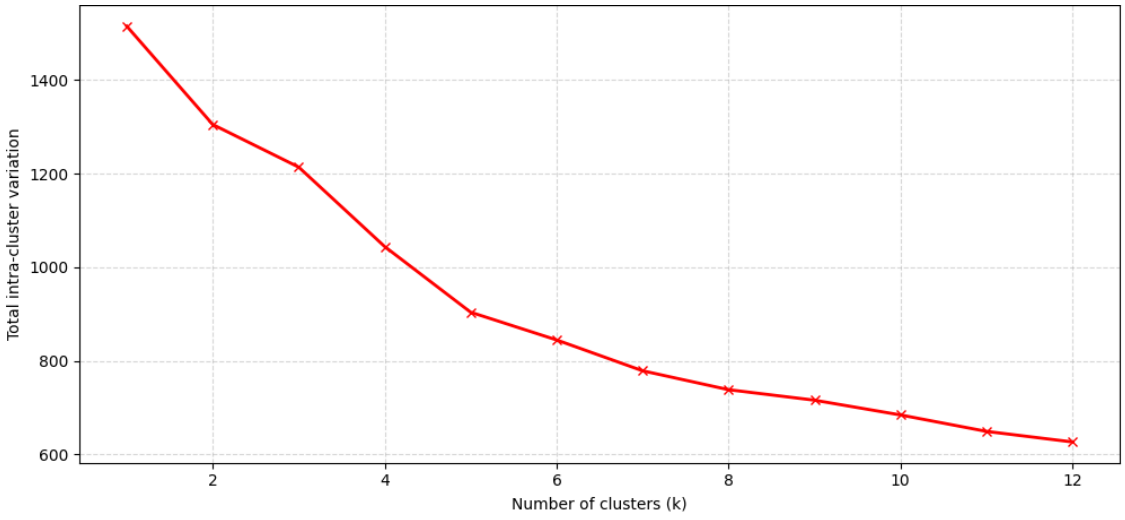

Figure 21. Elbow graph obtained using cosine distance and complete linkage on standardized data and determination of the optimal number of clusters

The Elbow curve in Figure 21 shows that the SSE values exhibit a distinct break, particularly at k=5, and that the rate of decrease slows down significantly after this number of clusters. Therefore, the five-cluster solution is considered a suitable option for the standardized data. Figure 22 presents the Elbow graph obtained using Euclidean distance and complete linkage on the normalized data, along with the determination of the optimal number of clusters.

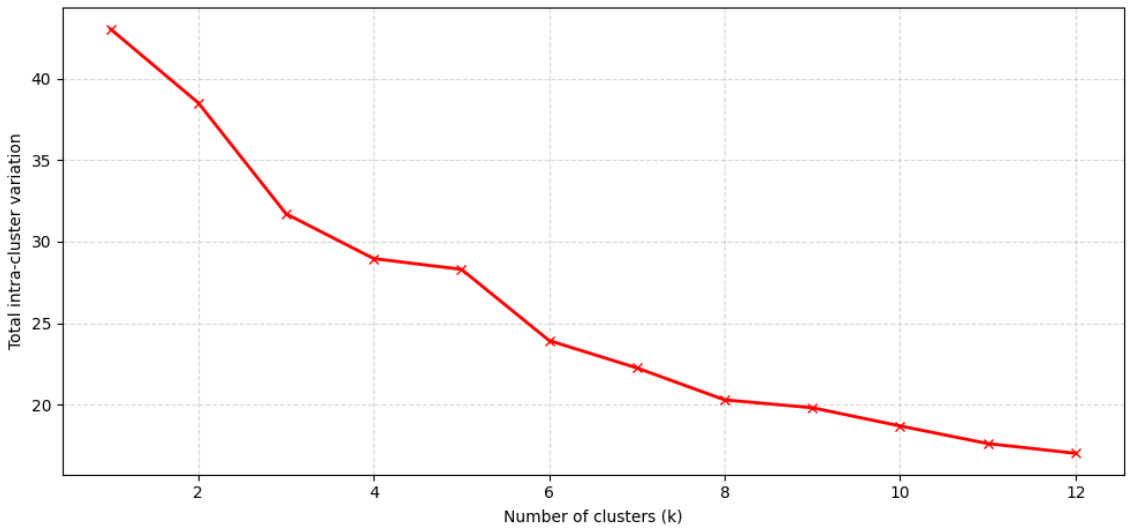

Figure 22. Elbow graph obtained using Euclidean distance and complete linkage on normalized data and determination of the optimal number of clusters

In the Elbow graph in Figure 22, a distinct break is observed, particularly at the k=3 point. This finding indicates that a three-cluster solution in the normalized data can meaningfully represent the data structure. Figure 23 shows the Elbow graph obtained using Euclidean distance and complete linkage on the standardized data, along with the determination of the possible optimal number of clusters.

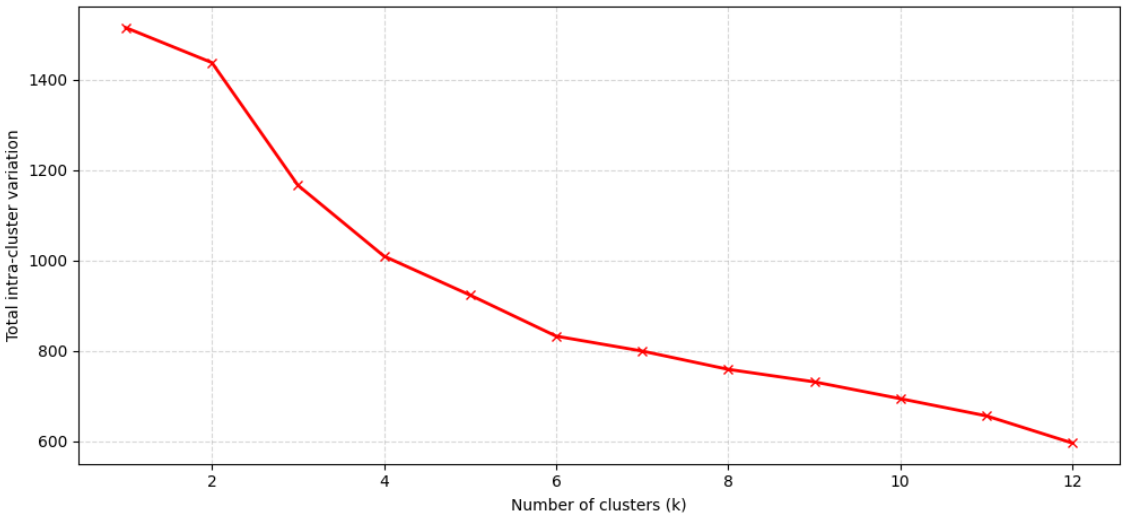

Figure 23. Elbow graph obtained using Euclidean distance and complete linkage on standardized data and determination of the possible optimal number of clusters

The curve in Figure 23 shows a distinct break, particularly in the range k=4-6) (it can be seen that the improvement effect of the clusters added after these cluster numbers on the SSE remains marginal.

b. Silhouette Method

The silhouette method is a highly useful technique, particularly when distance measures defined on a ratio scale (e.g., Euclidean distance) are used and the goal is to obtain compact, clearly separated clusters. This method relies on average distances, as in group average linkage, and is known to yield the best results when dealing with clusters with a spherical structure (Rousseeuw, 1987). Two basic pieces of information are required to calculate silhouette values:

- The clustering structure obtained as a result of the selected clustering algorithm,
- The set of distances between all objects.

The silhouette value $s(i)$ calculated for each object allows for the visualization of inter-cluster proximity. This calculation is performed using the steps shown in Equation (8):

- Any object i in the data set is considered together with the set A to which it belongs. If set A contains other objects besides object i:

a(i) = is defined as the average distance of object i from all other objects in its own set A. This measures intra-set similarity.

- Then, any set C other than the set containing object i is selected, and the following calculation is performed:

d(i,C) = the average distance of object *i* from all objects in set *C*

- After the *d(i,C)* values are calculated for all sets, the smallest value is selected and defined in Equation (8).

$$b(i) = \min_{C \neq A} d(i, C) \tag{8}$$

$b(i)$ in Equation (8) indicates the average distance to the nearest cluster outside the cluster to which the numbered observation belongs. This cluster is referred to as the "neighbor cluster" for the relevant observation.

- Finally, the silhouette value of object i is defined as shown in Equation (9) (Rousseeuw, 1987).

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{9}$$

According to the formula in Equation (9):

- $s(i) \approx 1$: The object is well grouped within its set.
- $s(i) \approx 0$: The object is on an uncertain border between two sets.
- $s(i) \approx -1$: The object may have been assigned to the wrong cluster.

When cluster A contains only a single observation, $a(i)$ cannot be defined) (therefore, $s(i) = 0$ is assumed for that observation. This assumption is considered the most appropriate approach methodologically, as it provides a neutral contribution to the silhouette value. Figure 24 shows the change in the average Silhouette coefficient for the cosine distance metric in normalized data according to the number of clusters.

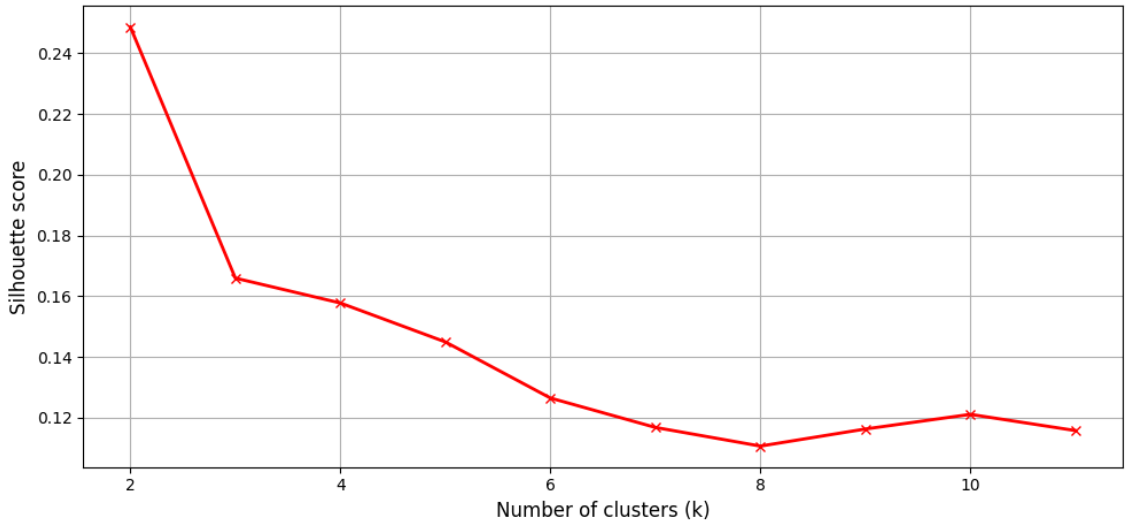

Figure 24. Change in the average Silhouette coefficient for the cosine distance metric in normalized data according to the number of clusters

Figure 24 shows that the average Silhouette score reaches its highest value of approximately 0.25 for $k = 2$ and gradually decreases for $k \geq 3$. This pattern suggests that the cosine distance and normalized data provide the best relative separation for the two-cluster solution. Figure 25 shows the change in the average Silhouette coefficient for the Euclidean distance metric in standardized data according to the number of clusters.

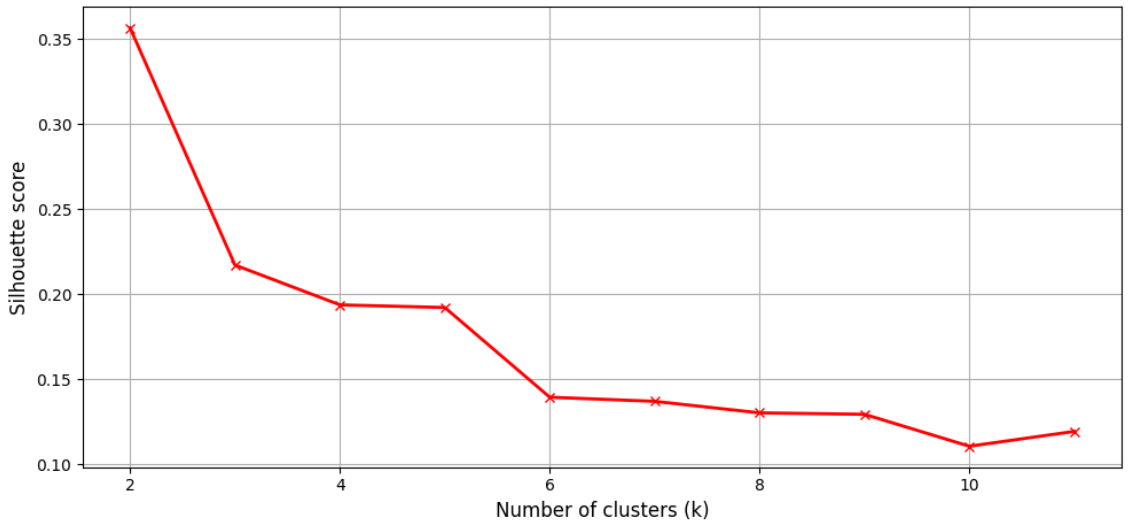

Figure 25. Change in the average Silhouette coefficient for the Euclidean distance metric in standardized data according to the number of clusters

Figure 25 shows that the average Silhouette score reaches its highest value of approximately 0.35 for $k = 2$, while it decreases significantly for $k = 3$. This finding indicates that the two-cluster structure is the most dominant and discriminative solution under the Euclidean distance metric in the standardized data. Figure 26 shows the change in the average Silhouette coefficient for the Euclidean distance metric in normalized data according to the number of clusters.

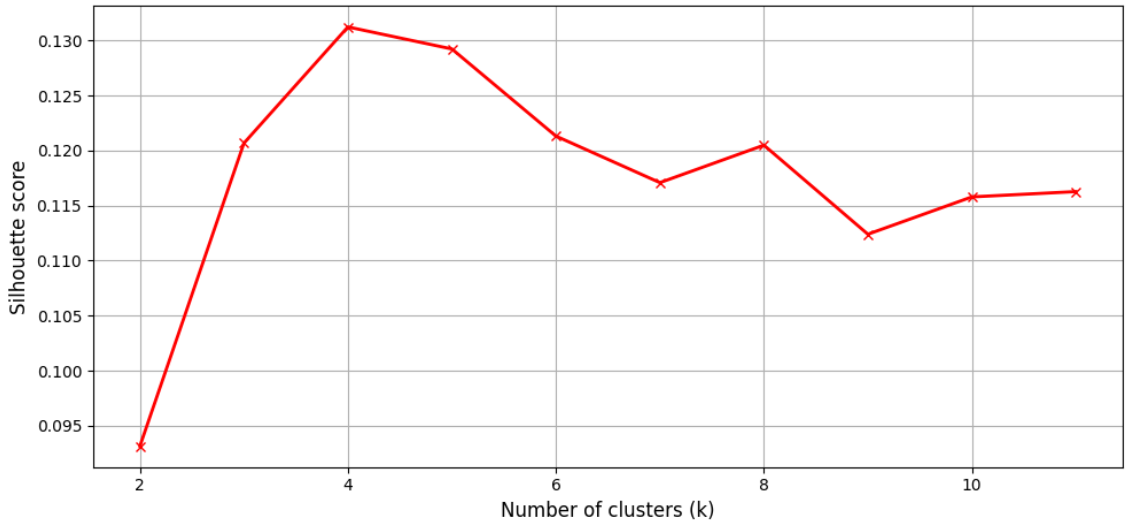

Figure 26. Change in the average Silhouette coefficient for the Euclidean distance metric in normalized data according to the number of clusters

Figure 26 shows that the average Silhouette score reaches its highest value at $k = 4$ (≈0.131), remains at a similar level for $k = 5$, and tends to decrease for $k \geq 6$. This pattern indicates that the four-cluster solution provides a more balanced and preferable separation under the Euclidean distance metric in the normalized data. Figure 27 shows the change in the average Silhouette coefficient according to the number of clusters for the cosine distance metric in the standardized data.

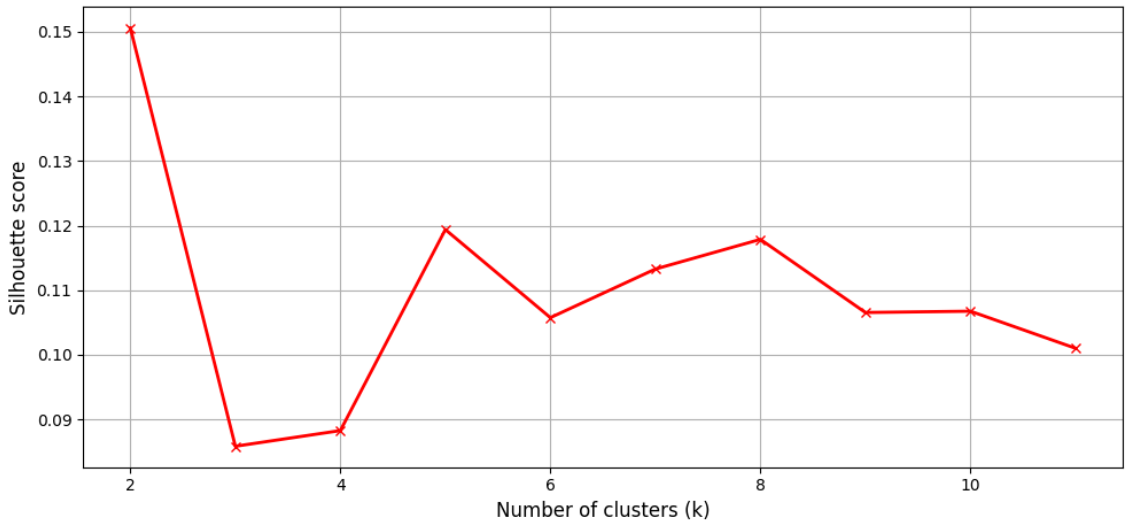

Figure 27. Change in the average Silhouette coefficient for the cosine distance metric in standardized data according to the number of clusters

Figure 27 shows that the average Silhouette coefficient again reaches its highest value for k = 2 (≈0.15), while the local peaks around k = 5 and k = 8 remain low and unstable. This situation indicates that the two-cluster solution is relatively more dominant under the cosine distance in the standardized data. However, when the Elbow and Silhouette results are evaluated together with hierarchical clustering dendrograms, it does not seem possible to determine a single and indisputable number of clusters for the data set. Different distance

measures (Euclidean, cosine) and scaling approaches (standardization, normalization) suggest different numbers of clusters, indicating that the decision should not be based on a single measure alone. Although both measures highlight a two-cluster structure when using cosine distance, the dendrograms show that the four-cluster solution offers a more consistent separation, with relatively low inter-cluster distances in the final merging steps) (therefore, relying solely on the Silhouette score to prefer the two-cluster solution may be misleading.

From the Euclidean distance perspective, dendrograms indicate the presence of a greater number of smaller subclusters, particularly those that may be clinically meaningful, in both normalized and standardized data. Although the Silhouette values are relatively low, the four-cluster solution is considered to better reflect the dendrogram structure and represent the heterogeneity in the dataset in a more clinically explanatory manner. Therefore, considering the findings obtained from different distance measures and scaling strategies together, the four-cluster solution was adopted as the baseline scenario in the continuation of the study.

IV. K-means Clustering

The k-means algorithm is one of the most commonly used clustering methods in data mining and is widely preferred, especially for analyzing large datasets. First proposed by MacQueen in 1967, this algorithm is considered one of the simplest yet effective partition-based approaches in the field of unsupervised learning. The algorithm divides data objects into k clusters) (it aims to converge to a local minimum by iteratively updating cluster centers and cluster memberships. Thus, the goal is for clusters to be relatively more compact within themselves and more distinct between clusters (Na et al, 2010).

Among non-hierarchical clustering methods, the k-means algorithm is considered one of the fundamental approaches that stands out due to its simple structure and wide range of applications (Ünal et al, 2011). Widely used in many different disciplines such as health, bioinformatics, marketing, and social sciences, the k-means algorithm is particularly prominent for its ability to discover patterns and distinguish groups in multidimensional data sets. The objective function of the k-means clustering algorithm is shown in Equation (10) (Alsabti et al, 1997).

$$E = \sum_{j=1}^{k} \sum_{i_l} \left| i_l - w_j \right|^2 \tag{10}$$

In Equation (10):

- k = the number of clusters,
- i_l = the l. data point belonging to the i. cluster,
- w_j = the center of the j. cluster,
- $\left| i_l - w_j \right|^2$ = the square of the Euclidean distance between the data point and the cluster center,
- E = represents the total error or total clustering cost for all clusters.

This objective function aims to maximize intra-cluster homogeneity by minimizing the distance of each data point to its cluster center. In other words, a smaller E value indicates that the clusters are more compact and better separated from each other. During algorithm execution, cluster centers (centroids) are iteratively updated, and this error function is minimized step by step.

The iterative update relations used in the K-means algorithm are defined as shown in Equation (11) and Equation (12) (Alsabti et al, 1997).

- In Equation (11), each data point i_l is assigned to the nearest center w_j:

$$C_j = \{i_l : \left\|i_l - w_j\right\|^2 \leq \|i_l - w_m\|^2, \forall m, 1 \leq m \leq k\} \quad (11)$$

- In Equation (12), the center of each cluster is updated with the average of the points assigned to it:

$$w_j^{(t+1)} = \frac{1}{|C_j|} \sum_{i_l \in C_j} i_l \quad (12)$$

The iterative process of the K-means algorithm proceeds by assigning data points to the nearest cluster center at each step and then updating each cluster center (w_j) by taking the arithmetic mean of the observations belonging to that cluster. This process is repeated until the objective function E approaches a minimum) (thus, the clusters gradually acquire a more compact and separable structure.

In this study, two different configurations were evaluated to examine the effect of data preprocessing strategies and distance metrics on clustering performance:

- Euclidean distance on standardized data,
- cosine distance on normalized data.

While Euclidean distance directly reflects the absolute differences between variables, it can be negatively affected by scale effects when variables are defined on different scales. Therefore, the variables were standardized so that their mean is 0 and variance is 1) (then, the Euclidean metric was used to ensure that the actual geometric differences in the data space are represented more accurately.

In normalized data (reduced to the 0–1 range), the direction and trend of the variables are more prominent than their absolute magnitude. Since the cosine distance provides a magnitude-independent similarity measure based on the angular similarity between vectors, it allows patients with different absolute levels of clinical parameters but similar patterns to be grouped together.

The combined use of these two approaches has revealed both scale-dependent and scale-independent patterns in the dataset and allowed for a more comprehensive, multidimensional evaluation of the clustering results.

Result and Discussion

This section presents the results obtained with the proposed clustering approach on patient data and discusses the performance of the approaches. In the analysis, the performance of k-means learners based on Euclidean and cosine distances is compared with different scaling options, such as standardization and normalization. The relationship between the resulting clusters and key clinical variables and the risk profiles of the participants is shown. Figure 28 presents the prominent binary feature spaces of the regional k-means clustering results using Euclidean distance on standardized data. Figure 29 presents the total k-means clustering results using cosine distance on normalized data in the classified binary feature spaces.

Figure 28. Distribution of Euclidean distance based k-means clustering results in selected binary feature spaces on standardized data.

Figure 29. Distribution of cosine distance based k-means clustering results in selected binary feature spaces on normalized data.

The visualizations presented in Figures 28 and 29 demonstrate that the scaling and distance metric selections produce different cluster structures in the patient data. In the standardized data, clusters obtained using Euclidean distance exhibit more pronounced separation in terms of age, cholesterol, ST depression, and maximum heart rate variables, while using cosine distance in the normalized data yielded an alternative segmentation where clusters partially overlapped but individuals with similar clinical tendencies were clustered together. Taken together, these two approaches allow for a more holistic assessment of patient groups in terms of both absolute risk and clinical patterns. Table 2 presents the mean values of clusters according to key clinical variables in the k-means cluster analysis based on Z-score standardization and Euclidean distance.

Table 2. Mean values of clusters according to basic clinical variables in k-means cluster analysis based on Z-score standardization and Euclidean distance

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
age	59,980	45,866	59,670	58,642
trestbps	141,660	123,036	141,352	124,528
chol	258,300	230,214	275,659	222,472
thalach	135,780	164,732	157,057	118,321
oldpeak	2,928	0,498	0,608	1,119

The cluster centers obtained after Z-score standardization, presented in Table 2, clearly demonstrate the absolute level differences between the patient groups. In terms of age, Cluster 2 represents a younger group with a mean age of 45.9 years, while the other clusters exhibit more advanced age profiles, ranging from 58 to 60 years. Resting blood pressure values indicate a pronounced hypertensive pattern in Cluster 1 (141.7 mmHg) and Cluster 3 (141.4 mmHg), while they are lower in Cluster 2 and Cluster 4. Similarly, serum cholesterol levels were highest in Cluster 3 (275.7 mg/dL) and lowest in Cluster 4 (222.5 mg/dL). Cluster 2 (164.7 beats/min) reflects relatively better cardiac capacity in terms of maximum heart rate, while Cluster 4 (118.3 beats/min) indicates a significant limitation in physical capacity. The highest mean value for ST-segment depression (oldpeak) was found in Cluster 1 (2.93), suggesting a more significant myocardial ischemic burden during stress testing in this group.

These findings indicate that the k-means clustering analysis revealed distinct subgroups based specifically on differences in hypertension, hypercholesterolemia, and exercise capacity) (Clusters 1 and 3 generally correspond to a higher cardiovascular risk profile, while Cluster 2 corresponds to a relatively lower-risk profile with younger age and better functional capacity. Table 3 presents the mean values of the clusters according to key clinical variables in the k-means clustering analysis based on min–max normalization and cosine distance.

Table 3. Mean values of clusters according to basic clinical variables in k-means clustering analysis based on min–max normalization and cosine distance

Variable	Cluster 1	Cluster 2	Cluster 3	Cluster 4
age	45,356	62,698	58,178	56,864
trestbps	123,752	118,283	143,567	139,203
chol	237,752	248,377	257,500	244,000
thalach	167,248	133,830	151,289	131,017
oldpeak	0,405	1,142	0,539	2,798

The cluster centers based on min–max normalization and cosine distance, presented in Table 3, group patient profiles based on pattern similarities rather than absolute levels. In terms of age, Cluster 1 corresponds to the younger age group (≈45 years), while Cluster 2 corresponds to the older age group (≈63 years). Resting blood pressure is lowest in Cluster 2 and highest in Cluster 3. Regarding cholesterol levels, Cluster 3 exhibits the highest levels, while Cluster 1 exhibits relatively lower levels. Maximum heart rate is highest in Cluster 1 and lowest in Cluster 4. In terms of ST depression (oldpeak), Cluster 4 exhibits significant stress burden, while Cluster 1 exhibits a low-risk profile.

Because cosine distance is sensitive to the directional patterns of variables, although partial overlap between clusters is observed, the variables age, thalach, and oldpeak appear to be particularly decisive in distinguishing between clusters. In general, the Z-score + Euclidean approach more clearly reveals differences in absolute risk levels, while the min–max + cosine approach offers a useful, complementary clustering perspective for integrating patient groups with similar clinical patterns.

In this study, the similarity between the clustering results obtained using cosine distance on normalized data and Euclidean distance on standardized data was calculated as Rand Index (RI) = 0.8179, indicating a highly consistent cluster structure across methods.

Conclusion

The conducted analyses allowed a comparative examination of k-means clustering based on different data transformation strategies (standardization, min–max normalization) and distance metrics (Euclidean, cosine) on clinical data related to heart disease. The similarity between the clustering results obtained using cosine distance for normalized data and Euclidean distance for standardized data was calculated as Rand Index (RI) = 0.8179, indicating a high level of cluster structure consistency across methods. This finding suggests that the internal cluster structure of the data set is relatively stable across both the scaling type and the distance metric, suggesting that the resulting clusters reflect meaningful segments based on the data structure rather than a random separation.

However, the fact that the RI value does not reach 1 also reveals the existence of certain structural differences arising from the distinction between standardization and normalization and Euclidean and cosine. Standardization compensates for variance differences between variables, making absolute level differences more pronounced) (whereas min-max normalization reduces values to the range [0,1] and focuses on relative magnitudes. Similarly, Euclidean distance emphasizes absolute level differences, while cosine distance is based on directional (pattern) similarity between observations. Therefore, some patient subgroups can be grouped differently based on "absolute risk level" under Euclidean distance and "clinical pattern similarity" under cosine distance. The resulting clusters revealed clinically significant subgroups that distinguish high- and low-risk profiles based on variables such as hypertension, hypercholesterolemia, exercise capacity, and ST depression.

These results demonstrate that in multidimensional and heterogeneous clinical datasets such as heart disease, clustering outcomes should be evaluated not only with

statistical fit criteria but also with regard to clinical interpretability and risk profiling. Even if high methodological similarity is achieved, it is crucial to thoroughly examine the clusters based on demographic, clinical, and biometric characteristics and to support them with clinician opinions to determine which clustering approach is more practical.

Future studies may include incorporating different unsupervised learning methods (e.g., Gaussian mixture models, DBSCAN, spectral clustering) into the comparative analysis, conducting external validation with larger sample and multicenter datasets, and correlating the resulting clusters with long-term clinical outcomes (mortality, rehospitalization, event frequency, etc.). Furthermore, integrating identified patient subgroups with risk prediction models and clinical decision support systems can be considered an important step toward increasing the usability of cluster-based profiling in clinical practice.

Author Contributions

Conceptualization: Akbas I., Taspinar Y.S. and Koklu M.) (Methodology: Akbas I., Taspinar Y.S. and Koklu M.) (Software: Akbas I.) (Validation: , Taspinar Y.S. and Koklu M. Formal analysis: Akbas I. and Taspinar Y.S.) (Investigation: Akbas I. and Koklu M.) (Resources: Akbas I. and Taspinar Y.S.) (Data curation: Akbas I.) (Writing original draft preparation: Akbas I. and Koklu M.) (Writing review and editing: Akbas I., Taspinar Y.S. and Koklu M.) (Visualization: Akbas I., Taspinar Y.S.) (Supervision: , Taspinar Y.S. and Koklu M.) (Project administration: Koklu M. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The dataset used in this study can be accessed via the link below: https://www.kaggle.com/datasets/kingabzpro/heart-disease-patients/data

Acknowledgments

The authors would like to express their gratitude to the Department of Computer Engineering, Faculty of Technology, Selcuk University, for providing access to their laboratory facilities used in this study.

Conflicts of Interest

The authors declare no conflict of interest.

Disclosure Statement

Generative artificial intelligence tools were employed for grammar refinement, linguistic clarity, and improvements in academic writing quality. These tools served as language-editing assistance within the manuscript preparation process.

References

Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. Technologies, 9(3), 52. https://doi.org/10.3390/technologies9030052

Alanazi, A. (2022). Using machine learning for healthcare challenges and opportunities. Informatics in Medicine Unlocked, 30, 100924. https://doi.org/10.1016/j.imu.2022.100924

Ali, L., Rahman, A., Khan, A., Zhou, M., Javeed, A., & Khan, J. A. (2019). An automated diagnostic system for heart disease prediction based on X^2 statistical model and optimally configured deep neural network. IEEE access, 7, 34938-34945. https://doi.org/10.1109/ACCESS.2019.2904800.

Ali, P. J. M. (2022). Investigating the Impact of min-max data normalization on the regression performance of K-nearest neighbor with different similarity measurements. ARO-The Scientific Journal of Koya University, 10(1), 85-91. https://doi.org/10.14500/aro.10955

Alsabti, K., Ranka, S., & Singh, V. (1997). An efficient k-means clustering algorithm (43). https://surface.syr.edu/eecs/43

Ambrish, G., Ganesh, B., Ganesh, A., Srinivas, C., & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings, 3(1), 127-130. https://doi.org/10.1016/j.gltp.2022.04.008

Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience, 2021(1), 8387680. https://doi.org/10.1155/2021/8387680

Capotosto, L., Massoni, F., De Sio, S., Ricci, S., & Vitarelli, A. (2018). Early diagnosis of cardiovascular diseases in workers: role of standard and advanced echocardiography. BioMed Research International, 2018(1), 7354691. https://doi.org/10.1155/2018/7354691

Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. (2022). An artificial intelligence model for heart disease detection using machine learning algorithms. Healthcare Analytics, 2, 100016. https://doi.org/10.1016/j.health.2022.100016

Chew, E. Y., Burns, S. A., Abraham, A. G., Bakhoum, M. F., Beckman, J. A., Chui, T. Y., Finger, R. P., Frangi, A. F., Gottesman, R. F., & Grant, M. B. (2025). Standardization and clinical applications of retinal imaging biomarkers for cardiovascular disease: a Roadmap from an NHLBI workshop. Nature Reviews Cardiology, 22(1), 47-63. https://doi.org/10.1038/s41569-024-01060-8

Cinar, I., Taspinar, Y. S., Kursun, R., & Koklu, M. (2022). Identification of corneal ulcers with pre-trained AlexNet based on transfer learning 2022 11th Mediterranean conference on embedded computing (MECO),

DeGuire, J., Clarke, J., Rouleau, K., Roy, J., & Bushnik, T. (2019). Blood pressure and hypertension. Health Rep, 30(2), 14-21. https://doi.org/10.25318/82-003-x201900200002

Erdem, K., Yasin, E., Yıldız, M. B., & Koklu, M. (2024). Classification of Heart Diseases with Ensemble Learning Algorithms. Sinop Üniversitesi Fen Bilimleri Dergisi, 9(2), 369-387. https://doi.org/10.33484/sinopfbd.1458580

Erdem, K., Yıldız, M. B., Yasin, E. T., & Köklü, M. (2023). A Detailed analysis of detecting heart diseases using artificial intelligence methods. Intelligent Methods In Engineering Sciences, 2(4), 115-124. https://doi.org/10.58190/imiens.2023.71

García-Vicente, C., Chushig-Muzo, D., Mora-Jiménez, I., Fabelo, H., Gram, I. T., Løchen, M.-L., Granja, C., & Soguero-Ruiz, C. (2023). Evaluation of synthetic categorical data generation techniques for predicting cardiovascular diseases and post-hoc interpretability of the risk factors. Applied Sciences, 13(7), 4119. https://doi.org/10.3390/app13074119

Gaziano, T., Reddy, K. S., Paccaud, F., Horton, S., & Chaturvedi, V. (2006). Cardiovascular disease. In D. Jamison, J. Breman, & A. Measham (Eds.), Disease Control Priorities in Developing Countries. 2nd edition (2nd ed.). The International Bank for Reconstruction and Development / The World Bank.

Ghiasi, M. M., Zendehboudi, S., & Mohsenipour, A. A. (2020). Decision tree-based diagnosis of coronary artery disease: CART model. Computer methods and programs in biomedicine, 192, 105400. https://doi.org/10.1016/j.cmpb.2020.105400

Gorenoi, V., Schönermark, M. P., & Hagen, A. (2012). CT coronary angiography vs. invasive coronary angiography in CHD. GMS health technology assessment, 8, Doc02. https://doi.org/10.3205/hta000100

Habehh, H., & Gohel, S. (2021). Machine learning in healthcare. Current genomics, 22(4), 291-300. https://doi.org/10.2174/1389202922666210705124359

Haq, A. U., Li, J. P., Khan, J., Memon, M. H., Nazir, S., Ahmad, S., Khan, G. A., & Ali, A. (2020). Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data. Sensors, 20(9), 2649. https://doi.org/10.3390/s20092649

Hayta, E., Gencturk, B., Ergen, C., & Koklu, M. (2023). Predicting future demand analysis in the logistics sector using machine learning methods. Intelligent Methods In Engineering Sciences, 2(4), 102-114. https://doi.org/10.58190/imiens.2023.70

Jarman, A. M. (2020). Hierarchical cluster analysis: Comparison of single linkage, complete linkage, average linkage and centroid linkage method. Georgia Southern University, 29, 90240. https://doi.org/10.13140/RG.2.2.11388.90240

Kavitha, S., & Kaulgud, N. (2023). Quantum K-means clustering method for detecting heart disease using quantum circuit approach. Soft Computing, 27(18), 13255-13268. https://doi.org/10.1007/s00500-022-07200-x

Kim, S. (2015). ppcor: an R package for a fast calculation to semi-partial correlation coefficients. Communications for statistical applications and methods, 22(6), 665. https://doi.org/10.5351/CSAM.2015.22.6.665

Koklu, M., & Sabancı, K. (2016). Estimation of credit card customers payment status by using kNN and MLP. International Journal of Intelligent Systems and Applications in Engineering, 4(Special Issue-1), 249-251. https://doi.org/10.18201/ijisae.281901

Krittanawong, C., Virk, H. U. H., Bangalore, S., Wang, Z., Johnson, K. W., Pinotti, R., Zhang, H., Kaplin, S., Narasimhan, B., & Kitai, T. (2020). Machine learning prediction in

cardiovascular diseases: a meta-analysis. Scientific reports, 10(1), 16057. https://doi.org/10.1038/s41598-020-72685-1

Marutho, D., Handaka, S. H., & Wijaya, E. (2018). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news 2018 international seminar on application for technology of information and communication,

Mooney, S. J., & Pejaver, V. (2018). Big data in public health: terminology, machine learning, and privacy. Annual review of public health, 39(1), 95-112. https://doi.org/10.1146/annurev-publhealth-040617-014208

Morgenstern, J. D., Buajitti, E., O'Neill, M., Piggott, T., Goel, V., Fridman, D., Kornas, K., & Rosella, L. C. (2020). Predicting population health with machine learning: a scoping review. BMJ open, 10(10), e037860. https://doi.org/10.1136/bmjopen-2020-037860

Muthumani, N., & Akilandeswari, K. (2024). Optimized Feature Selection and Classification Framework for Cardiovascular Disease Using Statistical Normalization and Bio-Inspired Algorithms 2024 International Conference on Communication, Control, and Intelligent Systems (CCIS),

Na, S., Xumin, L., & Yong, G. (2010). Research on k-means clustering algorithm: An improved k-means clustering algorithm 2010 Third International Symposium on intelligent information technology and security informatics,

Nabel, E. G. (2003). Cardiovascular disease. New England Journal of Medicine, 349(1), 60-72. https://doi.org/10.1056/NEJMra035098

Nadeem, M. W., Goh, H. G., Khan, M. A., Hussain, M., & Mushtaq, M. F. (2021). Fusion-Based Machine Learning Architecture for Heart Disease Prediction. Computers, Materials and Continua, 67(2), 2481-2496. https://doi.org/10.32604/cmc.2021.014649

Ni, Z., Liu, K., & Kang, G. (2018). Research on cardiovascular disease prediction based on distance metric learning Journal of Physics: Conference Series,

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. arXiv preprint arXiv:1503.06462. https://doi.org/10.48550/arXiv.1503.06462

Popp, R. L. (1976). Echocardiographic assessment of cardiac disease. Circulation, 54(4), 538-552. https://doi.org/10.1161/01.CIR.54.4.538

Prabhakaran, D., Jeemon, P., & Roy, A. (2016). Cardiovascular diseases in India: current epidemiology and future directions. Circulation, 133(16), 1605-1620. https://doi.org/10.1161/CIRCULATIONAHA.114.008729

Prasetyo, S. Y., Kurniawan, A., Sihotang, E. F. A., Puspita, R., & Setiawan, K. E. (2023). Heart disease risk prediction using K-nearest neighbor: A study of Euclidean and cosine distance metrics 2023 3rd International Conference on Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS),

Priyadarshinee, S., & Panda, M. (2022). Improving prediction of chronic heart failure using smote and machine learning 2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA),

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7

Saritas, M. M., Kursun, R., & Koklu, M. (2025, 06–07 October 2025). Detection of Bone Fractures in X-ray Images with Machine Learning Methods Using InceptionV3 Deep Features 3rd International Conference on Pioneer and Innovative Studies (ICPIS 2025),

Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 345. https://doi.org/10.1007/s42979-020-00365-y

Sharma, S., & Batra, N. (2019). Comparative study of single linkage, complete linkage, and ward method of agglomerative clustering 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon),

Shilaskar, S., & Ghatol, A. (2013). Feature selection for medical diagnosis: Evaluation for cardiovascular diseases. Expert systems with applications, 40(10), 4146-4153. https://doi.org/10.1016/j.eswa.2013.01.032

Shrivastava, A., Chakkaravarthy, M., & Shah, M. A. (2023). A new machine learning method for predicting systolic and diastolic blood pressure using clinical characteristics. Healthcare Analytics, 4, 100219. https://doi.org/10.1016/j.health.2023.100219

Singh, A., & Kumar, R. (2020). Heart disease prediction using machine learning algorithms 2020 international conference on electrical and electronics engineering (ICE3),

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. University of Kansas Scientific Bulletin, 38(6), 1409–1438. https://sid.ir/paper/549615/en

Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. Digital health, 6, 2055207620914777. https://doi.org/10.1177/2055207620914777

Sumwiza, K., Twizere, C., Rushingabigwi, G., Bakunzibake, P., & Bamurigire, P. (2023). Enhanced cardiovascular disease prediction model using random forest algorithm. Informatics in Medicine Unlocked, 41, 101316. https://doi.org/10.1016/j.imu.2023.101316

Taspinar, Y. S., Cinar, I., & Koklu, M. (2022). Classification by a stacking model using CNN features for COVID-19 infection diagnosis. Journal of X-ray science and technology, 30(1), 73-88. https://doi.org/10.3233/XST-211031

Taspinar, Y. S., Cinar, I., Kursun, R., & Koklu, M. (2024). Monkeypox Skin Lesion Detection with Deep Learning Models and Development of Its Mobile Application. Public health, 500, 5.

Upadhyay, S., Dwivedi, A., Verma, A., & Tiwari, V. (2023). Heart disease prediction model using various supervised learning algorithm 2023 IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT),

Ünal, Y., Ekim, U., & Köklü, M. (2011). Unıversıte Ogrencılerın Ortak Zorunlu Derslerdekı Basarılarının K-Means Algorıtması Ile Incelenmesı. Engineering Sciences, 6(1), 342-347. https://doi.org/10.12739/nwsaes.v6i1.5000067037

Yasin, E., & Koklu, M. (2025). A comparative analysis of machine learning algorithms for waste classification: inceptionv3 and chi-square features. International Journal of Environmental Science and Technology, 22(10), 9415-9428. https://doi.org/10.1007/s13762-024-06233-z

Yim, O., & Ramdeen, K. T. (2015). Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. The quantitative methods for psychology, 11(1), 8-21. https://doi.org/10.20982/tqmp.11.1.p008

Zhao, D., Liu, J., Wang, M., Zhang, X., & Zhou, M. (2019). Epidemiology of cardiovascular disease in China: current features and implications. Nature Reviews Cardiology, 16(4), 203-212. https://doi.org/10.1038/s41569-018-0119-4

Zhou, H., Deng, Z., Xia, Y., & Fu, M. (2016). A new sampling method in particle filter based on Pearson correlation coefficient. Neurocomputing, 216, 208-215. https://doi.org/10.1016/j.neucom.2016.07.036