**Scientific Journal of Informatics**
Vol. 11, No. 4, Nov 2024
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

# Heart Disease Clustering Modeling Using a Combination of the K-Means Clustering Algorithm and the Elbow Method

**Jihan Wala[1], Herman[2*], Rusydi Umar[3], Suwanti[4]**
[1, 2, 3] Master of Informatics, Universitas Ahmad Dahlan, Indonesia
[4] Bachelor of Informatics, Universitas Muhammadiyah Maumere, Indonesia

**Abstract.**
**Purpose:** This work seeks to create an efficient clustering model for categorising heart disease patient data with the K-Means algorithm while optimising the cluster count using the Elbow approach. The primary aim is to develop a precise and efficient model for detecting heart disease risk patterns, reducing underfitting and overfitting errors, and accelerating the processing of extensive datasets.
**Methods:** The model development employs a mix of the K-Means algorithm and the Elbow approach on a heart disease patient dataset sourced from the Kaggle Repository, including 303 patient data points. The study procedures entail: first, pre-processing to rectify missing values and standardise the data. The value of K is thereafter established for experimentation and the examination of clustering outcomes using K-Means. The Elbow approach is then used to determine the best number of clusters by computing the SSE (Sum of Squared Errors) and generating an elbow graph. The last phase involves interpreting the findings from each cluster derived from the optimum clustering.
**Result:** This study's findings demonstrate that the K-Means clustering algorithm, using the Elbow approach, effectively discovered two ideal clusters (k=2) within a dataset of 303 heart disease patients. The analysis of the first cluster indicates that, on average, individuals possess considerable risk factors for heart disease, while the second cluster consists of patients exhibiting signs of low risk for heart disease events.
**Novelty:** This work integrates the K-Means algorithm with the Elbow technique to ascertain the ideal number of clusters for classifying heart disease patients, therefore enhancing accuracy and mitigating the risks of overfitting and underfitting. Moreover, it may serve as a foundation for health policy, enhancing clinical decision-making and facilitating additional study into the use of machine learning technologies in the healthcare industry.

**Keywords**: K-means clustering, Elbow method, Heart disease, Disease severity, Patient emergencies
**Received** September 2024 / **Revised** October 2024 / **Accepted** November 2024



## INTRODUCTION

Heart disease, medically known as cardiovascular disease, is a medical condition that involves disorders of the heart or the blood vessels that supply blood to the heart [1], [2]. Patients suffering from heart disease often experience various complications in the body's metabolic cycle [3]. Some types of heart disease include coronary artery disease (CAD), heart attack (myocardial infarction), heart failure, arrhythmia, cardiomyopathy, congenital heart disease, and high blood pressure (hypertension) [4], [5]. The World Health Organization (WHO) has identified heart disease as the leading cause of death globally [6], accounting for approximately 30% of all annual deaths [7]. Each year, the number of deaths due to heart disease reaches approximately 12 million, with 80% of these deaths occurring as a direct result [8]. The prevalence of heart disease continues to increase in various countries, driven by several factors, including unhealthy lifestyles, genetic factors, and other medical complications [9], [10]. Additional factors that contribute to heart disease include obesity, high cholesterol, alcohol consumption, lack of physical activity, high blood pressure, diabetes, and stress [10]. A multifaced approach that incorporates lifestyle modifications, pharmacological interventions, and technological advancements holds promise for improving the quality of life for individuals diagnosed with heart disease. The management of heart disease is evolving, emphasizing prevention, early diagnosis, and more efficacious treatments. It is of the utmost importance to prevent heart disease, as it reduces the risk of developing the condition, enhances quality of life, alleviates the strain on the healthcare system. The advancement of technology, particularly machine learning, has the potential to significantly increase the prevention and management of heart disease.

*Corresponding author.
Email addresses: jihanwala4@gmail.com (wala), hermankaha@mti.uad.ac.id (herman)*, rusydi@mti.uad.ac.id (umar), suwantis009@gmail.com (suwanti)
DOI: 10.15294/sji.v11i4.14096

Machine learning is an essential instrument in cardiovascular prevention/ treatment, which solves many traditional problems. It can analyze vast, complex datasets such as medical records and lab reports + lifestyle data to find out what patterns relate to health. Sophisticated and targeted analytical techniques using this technology have the potential to mark a considerable improvement in heart disease prevenvionand treatment. Radhika et al., (2021) also applied several machine learning models such as k-Nearest Neighbours (KNN), Support Vector Machines (SVM): Naive Bayes, Logistic regression: Decision tree algorithm and random forest algorithm to classify different cardiac diseases in their study. KNN and random forest algorithms had the best results for accuracy, achieving an overall score of 88.52%, according to their findings [11]. Rahma et al. (2022) Cardiovascular disease Hu et al. The outcome of their analysis is in line where the naïve Bayes method displays 97% accuracy which triumphed over other machine learning algorithms themselves, included within the study [12]. Tick et al. (2021) used a multilayer perceptron (MLP)-Bp artificial neural network to predict myocardial infarction in patients with cardiovascular disease. These results also showed that the artificial neural networks could effectively classify heart disease patients with a learning rate of 0.25 and using only 25 neurons (Total accuracy was about:80.66% [13]. Also, machine learning provides clustering methods to classify heart disease instances. By focusing patients within healthcare systems based on profiles of diagnoses and symptoms, we can do a better job, including improving efficiency, managing resources more strategically to be able to expand programs that are individualized and targeted.

The k-means allows that data to be partitioned into categories. In the world of machine learning, we call this grouping, a fancier term for it is ‘clustering' [14]. In this approach, property data is divided into categories and separated then allocated to various disjoint clusters [15], [16]. In addition to being able to sort individuals according to health-related risk factors, these algorithms can also be used as a tool for the formulation of targeted prevention strategies per group increasing selection and intervention efficiency. The result shows that K-means clustering would discern undetectable health trends by simply using conventional method, thus enabling timely detection of the wrong, which could be potentially beneficial in monitoring health. Such strategy is utilized for heart patients and through attributes, we classify the patient based on whether it has severe condition or not. Then the Elbow method is used to find out and verify how many clusters are there in which outcome. Deciding the number of clusters is paramount that affects all characteristics in quality, interpretability and overall efficiency of cluster results as well as each makes a substantial effect on the entire process of cluster analysis. The Elbow method can assist us to determine an optimum K for a given dataset so far which we used Euclidean distance-based k-means clustering algorithm. Subject itself might have more than one correct number let along its vague definition. Selecting too few clusters can lead to overfitting (you are allowing the model to add noise data as useful clustering even if they do not really help) or underfitting, where in your clusters are prevented from many differentiating factors. The determination of an appropriate number of clusters ensures meaningful and relevant clustering, facilitates enhanced interpretation of results, and optimizes the utilization of computational resources. In practical applications, such as heart disease prevention, an optimal number of clusters enables the formulation of more effective preventive programs tailored to individual risk groups.

In a recent study (Ashari et al., 2023) used four clustering methods, namely the elbow, silhouette, Davidson-Bouldin, and Calinski-Harabasz, were employed in the k-means algorithm for the classification of flood-affected areas in Jakarta. Additionally, the rand index method was utilized as a means of evaluation. The findings that indicated the optimal solution consisted of three clusters, with a value of 1, and that a two-cluster solution, with a value of 0.9182, was also highly effective. The validation and assessment procedures demonstrated that the optimal grouping was achieved with three clusters. The distribution indicated that 75.4% of the regions exhibited low-risk characteristics, 21.1% demonstrated medium-risk characteristics, and 3.5% displayed high-risk characteristics [14]. In their 2019 research, Umargono et al. employed a k-means clustering analysis to map teacher data from state schools across all districts and cities within the Central Java Province. The optimal number of clusters was determined through the elbow method, while the initial centroids were established using the mean and median values. It was observed that establishing initial cluster centers based on the mean data resulted in a 22.58% reduction in the number of iterations required to achieve uniformity within the clusters compared to a randomly selected initial centroid.

Furthermore, the Elbow method has been demonstrated to reduce the number of required iterations by 25% when determining the optimal number of clusters compared to alternative cluster numbers [16]. Alamsyah et al. (2022) decided on the use of K-Means Recency Frequency Monetary (RFM) cluster algorithm in

conjunction with the Elbow method, to segment clients within retail businesses. A clustering algorithm was applied in this study, producing three clusters of consumers with the lowest possible Sum Of Square Error (SSE) score = 25.839,39 and Calinski-Harabasz Index value =36/625,89. The SSE and CHI were the highest, hence being chosen as cluster settings [17]. Any identification or clustering technique must be supported by the appropriate patient data and attribution techniques to achieve correct results (in principle). In prior investigation, the K-Means Clustering algorithm was validated to be effective in organizing patient data collected regarding heart disease by normalization and again fusion with an Elbow method on various datasets collecting symptoms of cardiac illness. In the case of heart disease research, major attributes include age, sex/chest pain, blood pressure/cholesterol/blood sugar and ECG. Results. As previously described [18], [19] each attribute reflects specific risk factors and health conditions experienced by patients. For instance, blood pressure and cholesterol are essential clues of heart disease risk; as well canary-in-the-coal-mine measurements like ECG results or maximum heart rate that might hint at how the ticker performs under stress. These characteristics are important as they help in personalizing diagnosis and treatment that will lead to a better next predictive algorithm.

By combining K-Means Clustering along with Elbow method, it helps to provide the accurate and data-driven clustering. Then resources can be allocated effectively, and patients will have better health near outcomes. It allows providers to identify patient subsets of varying severity and urgency, like patients at risk for death or other life-threatening situation where quick action is needed. Indonesia is a developing country with the same global burden of chronic and acute diseases including heart disease as elsewhere in Asia, whichis a great weight to the health system. Due to its high prevalence, in Indonesia too, there is value in classifying patients based on severity and urgency for the proper prioritization of care readying.

In view of the above provided background, this paper research classification by severity/urgency with heart disease patients through K-Means Grouping algorithm and Elbow Method. Related to our work of patient grouping using K-means clustering [4] in previous research study "Implementation of k-means clustering in heart disease patients" [5]. Conclusion This work helps improve the machine learning model of patient categorization for heart disease using K-Means and Elbow method to specify number clusters. This enables medical professionals to easily distinguish groups of patients at similar risk for a heart attack based on their individual biology, increasing the fidelity and quickening the diagnosis process. In addition, this research can decrease the amount of time and money spent on developing a good model that may improve its usefulness in real world healthcare.

## METHODS

### Research dataset

The data used in this research was secondary data obtained from the Kaggle website, especially the heart disease dataset by Abid Ali Awan. The selection of this dataset was based on several important reasons. First, the dataset comes from a well-known and trusted source within the data science community, ensuring its validity and reliability. Additionally, this dataset has been used in various previous studies, allowing for relevant comparisons with other studies [2], [3], [10]. Table 1 provides details of the heart disease dataset.

Table 1. Characteristics of the heart disease dataset

| No | Attribute | Description |
|---|---|---|
| 1 | *Age* | Patient age (years) |
| 2 | *Sex* | Patient gender |
| 3 | *Cp* | Type of chest pain |
| 4 | *Trestbps* | Resting blood pressure (mm Hg) |
| 5 | *Chol* | Serum cholesterol (mg/dl) |
| 6 | *Fbs* | Fasting blood sugar > 120 mg/dl |
| 7 | *Restecg* | Resting electrocardiographic results |
| 8 | *Thalach* | Maximum heart rate |
| 9 | *Exang* | Exercise-induced angina |
| 10 | *Oldpeak* | ST depression |
| 11 | *Slope* | Slope of the peak exercise ST segment |

Table 1 shows the detailed characteristics of the heart disease dataset. Each data point has 11 attributes. This research dataset consists of 303 records of heart disease patients. The Age attribute shows the patient's age range from 29 to 77 years. The Gender attribute indicates the gender of the patient, where a value of 1

represents male and a value of 0 represents female. The Cp (Chest Pain) attribute describes the type of chest pain experienced by the patient. It has four values: 1 for typical angina, which is classic pain due to narrowing of the coronary arteries, often indicating coronary artery heart disease; 2 for non-typical angina, which is chest pain that can feel sharp or burning, and although it doesn't follow a typical angina pattern, it can still indicate a heart problem. 3 for non-anginal pain is unlikely to be related to heart problems, could be caused by gastrointestinal or musculoskeletal issues or anxiety, and 4 for asymptomatic, where the individual experiences ischemia without pain, indicating the presence of severe heart disease but difficult to detect without routine examination. Trestbps (Resting Blood Pressure) shows the patient's blood pressure at rest, measured in mm Hg. Blood pressure < 120 mm Hg is considered low, 120 mm Hg is considered normal, and > 120 mm Hg is considered high. Chol (Cholesterol) represents the patient's blood cholesterol level in mg/dl, with levels < 140 mg/dl considered low, 140 mg/dl considered normal, and > 140 mg/dl considered high. FBS (Fasting Blood Sugar) shows the patient's fasting blood sugar level, with a value of 1 if fasting blood sugar exceeds 120 mg/dl and 0 if it is less than or equal to 120 mg/dl. Restecg (Resting Electrocardiography Results) shows the results of a resting electrocardiogram, with three values: 0 for normal, 1 for ST-T wave abnormalities, and 2 for left ventricular hypertrophy. Thalach (Maximum Achieved Heart Rate) represents the highest heart rate a patient achieves, with higher values indicating a greater risk of heart disease. Exang (Exercise-Induced Angina) indicates whether the patient experiences chest pain during exercise, with a value of 0 if there is no pain and 1 if there is pain. Oldpeak refers to the ST depression caused by exercise compared to when at rest; higher values signal a risk of heart disease, in individuals. Apart from oldpeak slope represents the inclination of the ST segment during peak exercise. Categorized into three types as 0 for slope 1 for slope and 2, for upward slope.

During this research projects initial phase of data preprocessing is essential to prepare the dataset for analysis purposes. This crucial step ensures that the dataset is devoid of any errors and ready, for examination. Data preparation primarily consists of rectifying missing values and standardizing the data. Missing values arise when certain attributes lack information in observations due, to potential data entry or collection related errors. It is imperative to detect and rectify these missing values effectively to prevent any outcomes that may result in conclusions. Furthermore, and importantly little information may adversely affect the datasets quality. Diminish the accuracy of the analysis performed on it. Having data without missing values makes it easier to interpret results and helps identify patterns or anomalies. After ensuring the completeness of data comes the process of normalizing it to standardize the value range, for all attributes in the dataset so that each attribute contributes equally to the analysis. This study employs a technique called feature scaling for normalization purposes which adjusts each attributes range to a scale, between 0 and 1. This approach facilitates comparison or integration of data from various sources or attributes with differing value ranges. The feature scaling formula is provided in equation (1) [20].

$$\mathrm{X_{new}} = \frac{X_{old}}{X_{max}} \quad (1)$$

In the feature scaling formula, $X_{new}$ represents the normalized attribute value, $X_{old}$ refers to the original value of the attribute to be normalized, and $X_{max}$ denotes the highest value across all data for the same attribute.

**Clustering model development**

The subsequent step is the development of a clustering model utilizing the K-Means Clustering method in conjunction with the Elbow technique. This approach aims to identify the optimal number of clusters, thereby enhancing the ability to isolate risk clusters among heart disease patients. Figure 1 illustrates the stages involved in constructing the k-means clustering model, integrated with the elbow method.

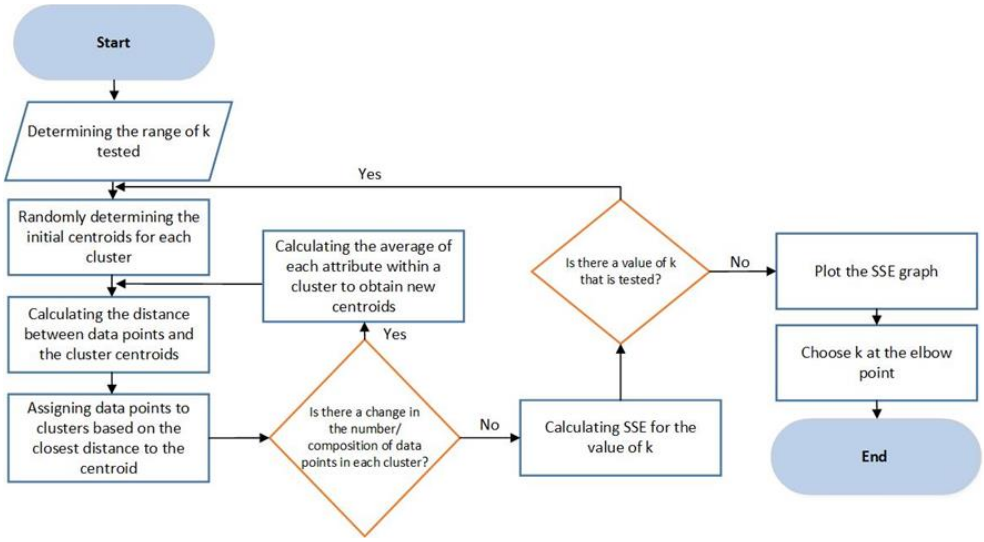


Figure 1. Flowchart of the k-means clustering model combined with the elbow method

The clustering model illustrated in Figure 1 starts by selecting the range of cluster values (k) to be tested. Following this, the k-means clustering algorithm is applied, with the goal of dividing the data points into the specified number of clusters. K-means groups data points into homogeneous clusters, where data in one cluster is more similar to each other than data in different clusters. In this algorithm the first step involves selecting the centroid for each cluster randomly without considering the distribution of the data points, in the dataset. As a result, every data point has an equal probability of being selected as the initial centroid. The next step involves assigning each data point (x) to a cluster by calculating the square root of the squared distance between each data point and its corresponding centroid (c) for every attribute. The Euclidean Distance formula is used for this calculation, as shown in equation [21].

$$\begin{aligned} (x_i, c_j) &= \sqrt{\textstyle\sum_{i=1}^{n}(x_i - c_j)^2} \\ d\,(x_i, c_j) &= \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + (x_3 - c_3)^2 + \ldots + (x_n - c_n)^2} \end{aligned} \quad (2)$$

Calculation of the Euclidean distance between each data point and each cluster centroid is marked with $d\,(x_i, c_j)$, where $x_i$ is the data point for each i-th attribute ($i = 1,2,3, \ldots . n$) with $i$ representing the sequence of data point attributes and $n$ represents the number of attributes. Meanwhile, $c_j$ is the centroid of each j-th attribute ($j = 1,2,3, \ldots . n$) with $j$ representing the sequence of centroid attributes in the cluster ($k$). Each row of data points and centroids calculates the distance to each of the identical or corresponding attributes. Based on the distance calculations, data points are assigned to the cluster nearest to the centroid. Once the clusters are formed, the average of each attribute within each cluster is computed to update the centroid's position. This process repeats iteratively until no significant changes occur in the cluster composition or the movement of data points. After all data points are correctly grouped into their respective clusters, the next step involves calculating the SSE (Sum of Squared Errors) value for each k value during the application of the elbow method.

The Elbow method helps identify the optimal number of clusters within the predefined range of cluster tests [22]. The optimal number of clusters is determined by calculating the SSE value, which is computed at the end of each iteration or once the data point composition in each cluster stabilizes. The formula for calculating SSE is presented in Equation 3 [16], [23].

$$\begin{aligned} SSE &= \sum_{j=1}^{k}\sum_{i=1}^{n}(x_i - c_j)^2 \\ SSE &= \textstyle\sum_{j=1}^{k}(x_1 - c_1)^2 + (x_2 - c_2)^2 + (x_3 - c_3)^2 + \ldots + (x_n - c_k)^2 \end{aligned} \quad (3)$$

The SSE value is calculated for each cluster, where xix_ixi represents the data point for the i-th attribute (I = 1, 2, 3, ….. ni =1, 2, 3, ….. ni = 1, 2, 3, ….. n) with iii denoting the order of attribute data points and nnn

representing the total number of attributes. Meanwhile, cjc_jcj is the centroid for the j-th attribute (j = 1, 2, 3, ….. nj = 1, 2, 3, ….. nj = 1, 2, 3, ….. n) with jjj indicating the order of centroid attributes within the cluster (k). For every row of data points and centroids in a cluster, the Squared Error (SE) is computed as the squared difference between each pair of identical or corresponding attributes. The SE values for all data points within a cluster are then summed to obtain the SSE value.

This process is repeated across the entire predefined range of clusters (k). Once all SSE values have been calculated for each k, a graph of SSE versus k is created to identify elbow points. The optimal number of clusters can be determined based on the elbow points on the SSE graph, where the rate of decline in SSE becomes sharp, and the points after that experience a slow and stable decline. In the context of heart disease, SSE helps assess how well patients are stratified by severity or emergency. Good grouping will produce clusters with patients who have similar conditions and low SSE.

## RESULTS AND DISCUSSIONS

The raw dataset of heart disease patients before normalization obtained from the Kaggle repository has 11 attributes shown in Table 2.

Table 2. Dataset rows before normalization

| Age | Sex | Cp | Trestbps | Chol | Fbs | Restecg | Thalach | Exang | Oldpeak | Slope |
|---|---|---|---|---|---|---|---|---|---|---|
| 63 | 1 | 1 | 145 | 233 | 1 | 2 | 150 | 0 | 2.3 | 3 |
| 64 | 1 | 1 | 110 | 211 | 0 | 2 | 144 | 1 | 1.8 | 2 |
| 58 | 0 | 1 | 150 | 283 | 1 | 2 | 162 | 0 | 1 | 1 |
| 66 | 0 | 1 | 150 | 226 | 0 | 0 | 114 | 0 | 2.6 | 3 |
| 69 | 0 | 1 | 140 | 239 | 0 | 0 | 151 | 0 | 1.8 | 1 |
| 40 | 1 | 1 | 140 | 199 | 0 | 0 | 178 | 1 | 1.4 | 1 |
| 51 | 1 | 1 | 125 | 213 | 0 | 2 | 125 | 1 | 1.4 | 1 |
| 34 | 1 | 1 | 118 | 182 | 0 | 2 | 174 | 0 | 0 | 1 |
| 52 | 1 | 1 | 118 | 186 | 0 | 2 | 190 | 0 | 0 | 2 |
| 65 | 1 | 1 | 138 | 282 | 1 | 2 | 174 | 0 | 1.4 | 2 |
| 59 | 1 | 1 | 170 | 288 | 0 | 2 | 159 | 0 | 0.2 | 2 |
| 52 | 1 | 1 | 152 | 298 | 1 | 0 | 178 | 0 | 1.2 | 2 |
| … | … | … | … | … | … | … | … | … | … | … |
| 57 | 0 | 4 | 140 | 241 | 0 | 0 | 123 | 1 | 0.2 | 2 |
| 68 | 1 | 4 | 144 | 193 | 1 | 0 | 141 | 0 | 3.4 | 2 |
| 57 | 1 | 4 | 130 | 131 | 0 | 0 | 115 | 1 | 1.2 | 2 |

The dataset in Table 2 consists of 303 data points sorted based on the CP (Chest Pain) attribute. This ordering was performed after consultation with medical experts, who identified CP as a leading indicator of potential heart and circulatory system problems.

Before normalization, missing values were handled, resulting in no missing values in the dataset. After that, data normalization was carried out using feature scaling with Equation (1). The normalized dataset is presented in Table 3.

Table 3. Dataset after normalization

| Age | Sex | Cp | Trestbps | Chol | Fbs | Restecg | Thalach | Exang | Oldpeak | Slope |
|---|---|---|---|---|---|---|---|---|---|---|
| 0.818 | 1 | 0.25 | 0.725 | 0.413 | 1 | 1 | 0.743 | 0 | 0.371 | 1.000 |
| 0.831 | 1 | 0.25 | 0.550 | 0.374 | 0 | 1 | 0.713 | 1 | 0.290 | 0.667 |
| 0.753 | 0 | 0.25 | 0.750 | 0.502 | 1 | 1 | 0.802 | 0 | 0.161 | 0.333 |
| 0.857 | 0 | 0.25 | 0.750 | 0.401 | 0 | 0 | 0.564 | 0 | 0.419 | 1.000 |
| 0.896 | 0 | 0.25 | 0.700 | 0.424 | 0 | 0 | 0.748 | 0 | 0.290 | 0.333 |
| 0.519 | 1 | 0.25 | 0.700 | 0.353 | 0 | 0 | 0.881 | 1 | 0.226 | 0.333 |
| 0.662 | 1 | 0.25 | 0.625 | 0.378 | 0 | 1 | 0.619 | 1 | 0.226 | 0.333 |
| 0.442 | 1 | 0.25 | 0.590 | 0.323 | 0 | 1 | 0.861 | 0 | 0.000 | 0.333 |
| 0.675 | 1 | 0.25 | 0.590 | 0.330 | 0 | 1 | 0.941 | 0 | 0.000 | 0.667 |
| 0.844 | 1 | 0.25 | 0.690 | 0.500 | 1 | 1 | 0.861 | 0 | 0.226 | 0.667 |
| 0.766 | 1 | 0.25 | 0.850 | 0.511 | 0 | 1 | 0.787 | 0 | 0.032 | 0.667 |
| 0.675 | 1 | 0.25 | 0.760 | 0.528 | 1 | 0 | 0.881 | 0 | 0.194 | 0.667 |
| … | … | … | … | … | … | … | … | … | … | … |
| 0.740 | 0 | 1 | 0.700 | 0.427 | 0 | 0 | 0.609 | 1 | 0.032 | 0.667 |
| 0.883 | 1 | 1 | 0.720 | 0.342 | 1 | 0 | 0.698 | 0 | 0.548 | 0.667 |
| 0.740 | 1 | 1 | 0.650 | 0.232 | 0 | 0 | 0.569 | 1 | 0.194 | 0.667 |

Table 3 shows that the values in the dataset had been transformed using feature scaling so that they are in the range 0 to 1. The first stage in creating a clustering model is to determine the range of clusters to be tested, from k=1 to k=10. For k=1, there is only one cluster, and the k-means clustering process continues until the first iteration is complete. The following explanation of the k-means clustering algorithm using the elbow method will illustrate the process for k=2. In this case, the centroid (c1) represents the centre point of the cluster (k1), and the centroid (c2) represents the cluster (k2). The initial centroid for each cluster is presented in Table 4.

Table 4. Initial centroids

| Centroid | Age | Sex | Cp | Trestbps | Chol | Fbs | Restecg | Thalach | Exang | Oldpeak | Slope |
|---|---|---|---|---|---|---|---|---|---|---|---|
| c1 | 0.844 | 1 | 0.25 | 0.69 | 0.500 | 1 | 1 | 0.861 | 0 | 0.226 | 0.667 |
| c2 | 0.818 | 0 | 0.5 | 0.7 | 0.346 | 0 | 0 | 0.886 | 0 | 0 | 0.333 |

After determining the centroid of the cluster, the process continued by calculating the distance between each data point and the centroid using the Euclidean Distance Equation (Equation 2). Equation 2 calculates $d(1,1)$, the distance between the first data point attribute and the centroid of cluster 1, and $d(1,2)$, the distance between the first data point attribute and the centroid of cluster 2. The following is the process for calculating the distance between the first data point and the centroid.

$$d\left(x_i, c_j\right) = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + (x_3 - c_3)^2 + \ldots + (x_n - c_n)^2}$$

$$d(1,1) = \sqrt{\begin{array}{c}(0.818 - 0.844)^2 + (1 - 1)^2 + (0.25 - 0.25)^2 + (0.725 - 0.69)^2 + (0.413 - 0.500)^2 + (1 - 1)^2 + (1 - 1)^2 + \\ (0.743 - 0.861)^2 + (0 - 0)^2 + (0.371 - 0.226)^2 + (1 - 0.667)^2\end{array}}$$

$$d(1,1) = 0.66$$

$$d(1,2) = \sqrt{\begin{array}{c}(0.818 - 0.818)^2 + (1 - 0)^2 + (0.25 - 0.5)^2 + (0.725 - 0.7)^2 + (0.413 - 0.346)^2 + (1 - 0)^2 + (1 - 0)^2 + \\ (0.743 - 0.886)^2 + (0 - 0)^2 + (0.371 - 0)^2 + (1 - 0.333)^2\end{array}}$$

$$d(1,2) = 1.75$$

The Euclidean Distance calculation shows that the first data point is 0.66 away from the centroid (c1), which is closer than the distance from the centroid (c2). This shows that the first data point is included in cluster 1. Table 5 shows the results of iteration 1, where 127 data points were clustered into cluster 1, and the remaining 176 data points were clustered into cluster 2.

Table 5. Cluster results for iteration 1

| Id | Age | Sex | Cp | Trestbps | Chol | Fbs | Restecg | Thalach | Exang | Oldpeak | Slope | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 117 | 0.753 | 1 | 0.75 | 0.7 | 0.374 | 1 | 1 | 0.817 | 0 | 0 | 0.333 | 1 |
| 125 | 0.844 | 1 | 0.25 | 0.25 | 0.5 | 1 | 1 | 0.861 | 0 | 0.226 | 0.667 | 1 |
| 140 | 0.662 | 1 | 0.75 | 0.625 | 0.434 | 1 | 1 | 0.822 | 0 | 0.387 | 0.667 | 1 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |
| 262 | 0.753 | 0 | 0.5 | 0.68 | 0.566 | 1 | 1 | 0.752 | 0 | 0 | 0.333 | 1 |
| 214 | 0.857 | 0 | 1 | 0.89 | 0.404 | 1 | 0 | 0.817 | 1 | 0.161 | 0.667 | 2 |
| 6 | 0.727 | 1 | 0.5 | 0.6 | 0.418 | 0 | 0 | 0.881 | 0 | 0.129 | 0.333 | 2 |
| 14 | 0.571 | 1 | 0.5 | 0.6 | 0.466 | 0 | 0 | 0.856 | 0 | 0 | 0.333 | 2 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |
| 295 | 0.818 | 0 | 1 | 0.62 | 0.349 | 0 | 0 | 0.673 | 1 | 0 | 0.667 | 2 |

Before proceeding to iteration 2, new centroids were determined for both clusters (k1 and k2). Determining this new centroid for both clusters used the calculation of the average distance between all data point attributes and the centroid in each cluster. The results of calculating this average distance are shown in Table 6.

Table 6. Average distance between data points and centroids in iteration 1

| k | Age | Sex | Cp | Trestbps | Chol | Fbs | Restecg | Thalach | Exang | Oldpeak | Slope |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.748 | 0.704 | 0.731 | 0.717 | 0.467 | 0.963 | 1 | 0.716 | 0.444 | 0.210 | 0.617 |
| 2 | 0.688 | 0.354 | 0.667 | 0.652 | 0.424 | 0.042 | 0.120 | 0.775 | 0.188 | 0.099 | 0.490 |

The calculation for the second iteration used the same process as the first iteration, using the distance of each data point and the new centroid shown in Table 6. This process then continued to the third and fourth iterations, but there were still changes in the composition of the data in each cluster. This indicates that the data points are still moving clusters in response to adjustments to the centroid position. The k-means clustering process ended in the fifth iteration because the composition of the data in each cluster had not changed from the previous iteration. The clustering results in iteration 5 are presented in Table 7.

Table 7. Cluster results from iteration 5

| Id | Age | Sex | Cp | Trestbps | Chol | Fbs | Restecg | Thalach | Exang | Oldpeak | Slope | Cluster |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 117 | 0.753 | 1 | 0.75 | 0.7 | 0.374 | 1 | 1 | 0.817 | 0 | 0 | 0.333 | 1 |
| 125 | 0.844 | 1 | 0.25 | 0.25 | 0.5 | 1 | 1 | 0.861 | 0 | 0.226 | 0.667 | 1 |
| 140 | 0.662 | 1 | 0.75 | 0.625 | 0.434 | 1 | 1 | 0.822 | 0 | 0.387 | 0.667 | 1 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |
| 262 | 0.753 | 0 | 0.5 | 0.68 | 0.566 | 1 | 1 | 0.752 | 0 | 0 | 0.333 | 1 |
| 214 | 0.857 | 0 | 1 | 0.89 | 0.404 | 1 | 0 | 0.817 | 1 | 0.161 | 0.667 | 2 |
| 6 | 0.727 | 1 | 0.5 | 0.6 | 0.418 | 0 | 0 | 0.881 | 0 | 0.129 | 0.333 | 2 |
| 14 | 0.571 | 1 | 0.5 | 0.6 | 0.466 | 0 | 0 | 0.856 | 0 | 0 | 0.333 | 2 |
| … | … | … | … | … | … | … | … | … | … | … | … | … |
| 295 | 0.818 | 0 | 1 | 0.62 | 0.349 | 0 | 0 | 0.673 | 1 | 0 | 0.667 | 2 |

After the K-Means clustering process is complete, the next step is to calculate the Sum of Squared Errors (SSE) value for each k using Equation (2). SSE results for k=2 are presented in Table 8.

Table 8. Total SSE value k=2

| k | SSE |
|---|---|
| k1 | 117.04 |
| k2 | 105.92 |
| Total SSE | 222.95 |

Table 8 shows the total SSE value for k=2. The total SSE for cluster k1 is 117.04, and for cluster k2, it is 105.92, resulting in a total SSE value of 222.95. Table 9 presents the total SSE value for datasets grouped into 1 cluster (k=1) to 10 clusters (k=10).

Table 9. Total SSE value for each number of clusters (k=1 to k=10)

| Number of Clusters (k) | Number of Data Points per Cluster | | | | | | | | | | Total SSE Score |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k=1 | 303 | | | | | | | | | | 765.88 |
| k=2 | 151 | 152 | | | | | | | | | 222.95 |
| k=3 | 45 | 156 | 102 | | | | | | | | 203.10 |
| k=4 | 45 | 133 | 52 | 73 | | | | | | | 187.94 |
| k=5 | 37 | 93 | 51 | 46 | 76 | | | | | | 178.13 |
| k=6 | 29 | 93 | 51 | 48 | 66 | 16 | | | | | 158.16 |
| k=7 | 22 | 90 | 28 | 42 | 61 | 16 | 44 | | | | 125.45 |
| k=8 | 26 | 40 | 51 | 38 | 35 | 16 | 41 | 56 | | | 93.17 |
| k=9 | 26 | 40 | 1 | 51 | 30 | 16 | 41 | 56 | 42 | | 79.67 |
| k=10 | 22 | 40 | 2 | 35 | 49 | 16 | 39 | 56 | 35 | 9 | 73.39 |

Table 9 shows the Sum of Squared Errors (SSE) value for each number of clusters (k=1 to k=10) resulting from applying the k-means algorithm to a dataset containing 303 heart disease patients. As the number of clusters increases, the total SSE value decreases, reflecting the better grouping of data into more homogeneous clusters. At k=1, when all data is grouped into one cluster, the total SSE value is very high (765.88), with a total of 303 data points indicating significant variations within the cluster. When the number of clusters is increased to k=2, the SSE value drops drastically to 222.95, with cluster 1 totaling 151 and cluster 2 totaling 152 data points, indicating that the two clusters are starting to share data more efficiently. The decline in SSE values continues with increasing clusters but begins to show a slowdown at k=6 to k=7, where SSE values reach 158.16 and 125.45. This indicates that adding clusters after that point has a minor impact on reducing variation within the cluster. At k=10, the total SSE value reaches its lowest point, namely 73.39. Then, from the results of the total SSE value, the optimal number of clusters is determined using an elbow graphic illustration. The graph of the total SSE value from k=1 to k=10 is presented in Figure 2.

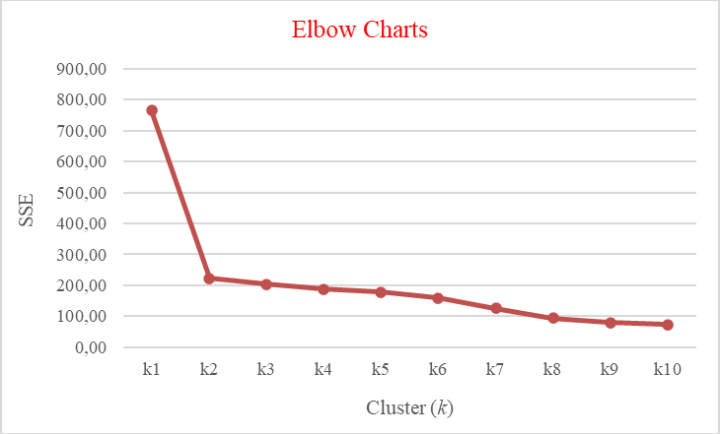


Figure 2. Elbow graph for each number of clusters k=1 to k=10

Figure 2 shows the Y axis representing Total SSE, which calculates the difference between the value for each datapoint attribute and the corresponding attribute value from the centroid. A smaller SSE value indicates better clustering because the data points are closer to the cluster centroid. Meanwhile, the X-axis shows the number of clusters (k) used in the k-means algorithm, ranging from k=1 to k=10. At k=1, the SSE is very high, approaching 800, because all data points are assigned to one cluster, resulting in high internal variation within the cluster. There is a significant decrease in the SSE value from point k=1 to point k=2. This sharp decrease indicates that increasing the number of clusters from 1 to 2 significantly improves data grouping because the SSE value is much smaller. After k=2, the SSE reduction becomes slower and more stable. This shows that each additional cluster beyond k=2 provides a minor increase in clustering. On the graph, there is a point that is often referred to as the 'elbow point,' namely k=2. This point represents the optimal balance between the number of clusters and SSE reduction. The results of clustering modelling using a combination of k-means clustering and the elbow plot method show that k=2 is the optimal cluster. The results of optimal cluster modelling k=2 in iteration 5 are presented in Table 10.

Table 10. Results of centroids in iteration 5

| k | Age | Sex | Cp | Trestbps | Chol | Fbs | Restecg | Thalach | Exang | Oldpeak | Slope | Number of Data Points |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.748 | 0.704 | 0.731 | 0.717 | 0.467 | 0.963 | 1 | 0.716 | 0.444 | 0.210 | 0.617 | 151 |
| 2 | 0.688 | 0.354 | 0.667 | 0.652 | 0.424 | 0.042 | 0.120 | 0.775 | 0.188 | 0.099 | 0.490 | 152 |

Table 10 shows the centroid data for each cluster (k) and the number of data points clustered into cluster k. Cluster 1 (k1) has 151 patients, and cluster 2 (k2) has 152 patients. The results in the table are still in the form of normalized data, so they need to be translated according to the initial data. This is done to find out the actual data pattern so that it is easy to carry out cluster analysis based on the severity or emergency level of heart disease patients. Below in Table 11, the results of the translation of initial data from two clusters (k=2) as a result of clustering modelling are presented.

Table 11. Data translation of optimal cluster results iteration 5 (k=2)

| k | Age | Sex | Cp | Trestbps | Chol | Fbs | Restecg | Thalach | Exang | Oldpeak | Slope |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 56 | 1 | 3 | 134 | 256 | 0 | 2 | 148 | 0 | 1.2 | 2 |
| 2 | 53 | 0 | 2 | 87 | 108 | 0 | 0 | 111 | 0 | 0.2 | 1 |

Analysis of clustering results showed different heart disease risk profiles for patients in each cluster. Cluster 1 consisted of older male patients, with a mean age of 56 years, who experienced Non-Anginal Pain (Type 3); this type of chest pain is less likely to be related to heart problems. Although there are no symptoms of chest pain, this cluster represents significant risk factors for heart disease, including high resting blood pressure (Trestbps=134 mm Hg), high cholesterol levels (Chol=256 mg/dL), and indications of left ventricular hypertrophy in electrocardiography (Restecg=2). A high oldpeak value (1.2) indicates possible exercise ischemia, which, coupled with ST-segment elevation, further emphasizes the potential for underlying cardiac disease. However, this cluster did not experience exercise-induced angina (Exang=0). The overall risk of heart disease is quite significant, so treatment is more aggressive and involves many medical and lifestyle aspects that are closely monitored.

In contrast, Cluster 2 mainly consisted of younger female patients, with a mean age of 53 years, who presented with Atypical Angina (Type 2) still related to the heart. However, the symptoms were not wholly consistent with typical angina. This cluster has low resting blood pressure (Trestbps=87 mm Hg) and deficient cholesterol levels (Chol=108 mg/dL), with average electrocardiography results (Restecg=0). Minimal ST segment depression during exercise (Oldpeak=0.2) and a flat ST segment slope indicate a low risk of heart disease. However, the maximum heart rate response to physical activity is lower (Thalach=111 bpm). The overall profile of this cluster indicates a low risk for heart disease events, so focus on prevention with long-term health maintenance and light monitoring.

The results of clustering modelling provide valuable insight into grouping heart disease patients based on the severity or emergency level of the condition through the combined application of the K-Means Clustering algorithm and the Elbow method in this research, the K-Means algorithm successfully segmented the patient population into distinct clusters, highlighting different levels of heart disease risk. Cluster 1 identified a group of older male patients who, despite not exhibiting symptoms of non-anginal chest pain, had high-risk factors for heart disease such as elevated blood pressure, high cholesterol, and notable electrocardiographic abnormalities, indicating a greater severity or need for urgent care. The elbow method was crucial in determining the optimal number of clusters, ensuring that these high-risk patients were grouped together based on their shared risk factors. On the other hand, Cluster 2 included younger female patients with atypical anginal chest pain and a lower risk profile for heart disease. The Elbow method's role in identifying two optimal clusters helps avoid over-segmentation, which could detract from focusing on clinically significant patient groups. By clustering patients more effectively, the combined use of K-Means Clustering and the Elbow method enhances the ability to pinpoint high-risk groups for closer monitoring and intervention, while also identifying low-risk groups that may require less intensive management. This integration of the Elbow method with K-Means clustering significantly improves the accuracy of patient grouping, particularly in distinguishing individuals based on the severity or urgency of their heart disease. This method not only supports the formulation of more tailored treatment strategies but also contributes to the efficient allocation of medical resources by prioritizing high-risk patients who need immediate care.

The current study on heart disease clustering, which utilized a combination of the K-Means Clustering Algorithm and the Elbow Method, demonstrates significant advancements over previous research. Unlike the clustering of non-communicable diseases in Banten [24], which primarily focused on geographic distribution using fixed k values in k-means clustering, this study emphasizes patient-based grouping based on the severity of heart disease. This patient-centric approach facilitates more personalized health interventions, although it may not offer the broader epidemiological insights provided by geographic clustering studies.

In contrast to the Enhanced Genetic Algorithm (EGA) and Fuzzy Weight-Based Support Vector Machine (FWSVM) methods used for early detection of heart disease [25], which prioritize classification accuracy, this study focuses on optimizing the number of clusters to enhance the precision of patient grouping. While both studies employed K-Means Clustering, the integration of the Elbow Method in our research introduces a powerful mechanism for determining the optimal number of clusters—a feature not explored in the EGA-FWSVM approach.

Additionally, this study offers a data-driven solution distinct from the randomized controlled trial of coronary heart disease interventions in families in India [26], which concentrated on lifestyle modification. The current research provides a technical solution that complements lifestyle interventions by effectively identifying high-risk patient clusters through appropriate clustering.

The main contribution of this research lies in integrating the K-Means Clustering Algorithm with the Elbow Method, producing optimal clustering results specifically for heart disease patients. This approach enhances the accuracy of patient grouping and holds significant potential for integration into health systems and improved resource management efficiency. By focusing on clustering based on disease severity, this study paves the way for future research to explore more sophisticated clustering algorithms and their applications in personalized medicine. Suggestions for further research include several strategic steps to deepen the analysis and increase the accuracy and relevance of grouping heart disease patients. First, additional clinical validation should be performed to ensure that the resulting clusters accurately reflect medical assessments

and provide practical guidance in patient care. This involves collaboration with medical experts to verify that the groupings are consistent with clinical diagnoses and treatment protocols.

Furthermore, exploring other clustering methods, such as hierarchical clustering or DBSCAN, can be applied to compare results and determine whether any additional structures can be identified that may not be revealed through K-Means Clustering. Analyzing additional attributes is also essential, including medical factors that may improve clustering accuracy and provide deeper insights into risk factors influencing heart health.

Finally, conducting longitudinal studies will allow for monitoring the effectiveness of treatment strategies implemented based on these groupings and evaluating their impact on patients' long-term cardiovascular health outcomes. This approach will provide a more comprehensive understanding of how appropriate grouping can influence disease management and overall health outcomes.

## CONCLUSION

This study demonstrates that the combination of the K-Means Clustering Algorithm and the Elbow Method can effectively classify heart disease patients based on their risk levels and the severity of emergency conditions. The optimal point, or 'elbow point,' is observed at k=2, where a balance between the number of clusters and reduced variability within them is achieved. The findings revealed that the two optimal clusters (k=2) separated patients into high and low-risk groups, each characterized by distinct clinical features. Patients in the high-risk cluster require more intensive medical intervention, while those in the low-risk group benefit from preventive measures and less intensive monitoring. The Elbow method efficiently identified the optimal cluster number, maximizing clustering effectiveness while minimizing within-cluster variability, as evidenced by a significant reduction in SSE values. The results of this research offer valuable insights into data-driven clustering for managing cardiac patients, and highlight the potential of incorporating this model into healthcare systems to enhance resource allocation and patient care. Moreover, this study paves the way for future research by encouraging the exploration of more advanced clustering techniques and additional medical attributes to improve clustering precision, along with clinical validation to ensure the applicability of the clustering outcomes in medical practice.

## ACKNOWLEDGEMENT

This research was supported by Directorate of Research, Technology, and Community Service Ministry of Education, Culture, Research and Technology, Indonesia under the Grant No. 107/E5/PG.02.00.PL/2024 0609.12/LL5-INT/AL.04/2024; 070/PTM/LPPM/UAD/VI/2024

## REFERENCES

[1] A. Anagnostopoulou, N. Eleftherakis, and E. Karanasios, “Obesity and Underweight in Children Who Undergo Catheterization for Congenital Heart Disease: A Retrospective Study,” *Glob. Pediatr.*, vol. 7, no. December, p. 100095, 2024, doi: 10.1016/j.gpeds.2023.100095.

[2] A. Singh and R. Kumar, “Heart Disease Prediction Using Machine Learning Algorithms,” *Int. Conf. Electr. Electron. Eng. ICE3 2020*, pp. 452–457, Feb. 2020, doi: 10.1109/ICE348803.2020.9122958.

[3] D. Shah, S. Patel, and S. K. Bharti, “Heart Disease Prediction using Machine Learning Techniques,” *SN Comput. Sci.*, vol. 1, no. 6, Nov. 2020, doi: 10.1007/s42979-020-00365-y.

[4] I. Tougui, A. Jilbab, and J. El Mhamdi, “Heart Disease Classification Using Data Mining Tools and Machine Learning Techniques,” *Health Technol. (Berl).*, vol. 10, no. 5, pp. 1137–1144, 2020, doi: 10.1007/s12553-020-00438-1.

[5] Pahwa Kanika and Kumar Ravinder, “Prediction of Heart Disease Using Hybrid Technique For Selecting Features,” *IEEE Uttar Pradesh Sect. Int. Conf. Electr. Comput. Electron.*, 2017, doi: 10.1109/UPCON.2017.8251100.

[6] G. Manikandan, B. Pragadeesh, V. Manojkumar, A. L. Karthikeyan, R. Manikandan, and A. H. Gandomi, “Classification Models Combined With Boruta Feature Selection For Heart Disease Prediction,” *Informatics Med. Unlocked*, vol. 44, Jan. 2024, doi: 10.1016/j.imu.2023.101442.

[7] D. Hassan, H. I. Hussein, and M. M. Hassan, “Heart Disease Prediction Based On Pre-Trained Deep Neural Networks Combined With Principal Component Analysis,” *Biomed. Signal Process. Control*, vol. 79, Jan. 2023, doi: 10.1016/j.bspc.2022.104019.

[8] R. Valarmathi and T. Sheela, “Heart Disease Prediction Using Hyper Parameter Optimization (HPO) Tuning,” *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, doi:

10.1016/j.bspc.2021.103033.
[9] I. D. Mienye, Y. Sun, and Z. Wang, “An Improved Ensemble Learning Approach For The Prediction Of Heart Disease Risk,” *Informatics Med. Unlocked*, vol. 20, Jan. 2020, doi: 10.1016/j.imu.2020.100402.
[10] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. W. Quinn, and M. A. Moni, “Heart Disease Prediction Using Supervised Machine Learning Algorithms: Performance Analysis And Comparison,” *Comput. Biol. Med.*, vol. 136, Sep. 2021, doi: 10.1016/j.compbiomed.2021.104672.
[11] R. Radhika and S. Thomas George, “HEART DISEASE CLASSIFICATION USING MACHINE LEARNING TECHNIQUES,” *J. Phys. Conf. Ser.*, vol. 1937, no. 1, p. 012047, Jun. 2021, doi: 10.1088/1742-6596/1937/1/012047.
[12] M. M. Rahma and A. D. Salman, “Heart Disease Classification–Based on the Best Machine Learning Model,” *Iraqi J. Sci.*, pp. 3966–3976, Sep. 2022, doi: 10.24996/ijs.2022.63.9.28.
[13] V. K. Tick, N. Y. Meeng, N. F. Mohammad, N. H. Harun, H. Alquran, and M. F. M. Mohsin, “Classification of Heart Disease using Artificial Neural Network,” *J. Phys. Conf. Ser.*, vol. 1997, no. 1, p. 012022, Aug. 2021, doi: 10.1088/1742-6596/1997/1/012022.
[14] I. F. Ashari, E. Dwi Nugroho, R. Baraku, I. Novri Yanda, and R. Liwardana, “Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta,” *J. Appl. Informatics Comput.*, vol. 7, no. 1, pp. 89–97, Jul. 2023, doi: 10.30871/jaic.v7i1.4947.
[15] O. S. Faragallah, H. M. El-Hoseny, and H. S. El-sayed, “Efficient brain tumor segmentation using OTSU and K-means clustering in homomorphic transform,” *Biomed. Signal Process. Control*, vol. 84, p. 104712, Jul. 2023, doi: 10.1016/j.bspc.2023.104712.
[16] E. Umargono, J. E. Suseno, and V. G. S. K., “K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median,” in *Proceedings of the International Conferences on Information System and Technology*, SCITEPRESS - Science and Technology Publications, 2019, pp. 234–240. doi: 10.5220/0009908402340240.
[17] A. Alamsyah *et al.*, “Customer Segmentation Using the Integration of the Recency Frequency Monetary Model and the K-Means Cluster Algorithm,” *Sci. J. Informatics*, vol. 9, no. 2, pp. 189–196, Nov. 2022, doi: 10.15294/sji.v9i2.39437.
[18] R. C. Ripan *et al.*, “A Data-Driven Heart Disease Prediction Model Through K-Means Clustering-Based Anomaly Detection,” *SN Comput. Sci.*, vol. 2, no. 2, p. 112, Apr. 2021, doi: 10.1007/s42979-021-00518-7.
[19] R. Indrakumari, T. Poongodi, and S. R. Jena, “Heart Disease Prediction using Exploratory Data Analysis,” *Procedia Comput. Sci.*, vol. 173, pp. 130–139, 2020, doi: 10.1016/j.procs.2020.06.017.
[20] A. Masitha, M. K. Biddinika, and H. Herman, “K Value Effect on Accuracy Using the K-NN for Heart Failure Dataset,” *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 22, no. 3, pp. 593–604, Jul. 2023, doi: 10.30812/matrik.v22i3.2984.
[21] B. Rizki, N. G. Ginasta, M. A. Tamrin, and A. Rahman, “Customer Loyality Segmentation on Point of Sale System Using Recency-Frequency-Monetary (RFM) and K-Means,” *J. Online Inform.*, pp. 130–136, Dec. 2020, doi: 10.15575/join.v5i2.511.
[22] S. Juanita and R. Cahyono, “K-MEANS CLUSTERING WITH COMPARISON OF ELBOW AND SILHOUETTE METHODS FOR MEDICINES CLUSTERING BASED ON USER REVIEWS,” *J. Tek. Inform.*, vol. 5, pp. 283–289, Feb. 2024, doi: 10.52436/1.jutif.2024.5.1.1349.
[23] N. González-Cabrera, J. Ortiz-Bejar, A. Zamora-Mendez, and M. R. Arrieta Paternina, “On the Improvement of representative demand curves via a hierarchical agglomerative clustering for power transmission network investment,” *Energy*, vol. 222, p. 119989, May 2021, doi: 10.1016/j.energy.2021.119989.
[24] T. A. Munandar and A. Yunizar Yusuf Pratama, “Regional Clustering Based on Types of Non-Communicable Diseases Using k-Means Algorithm,” *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 23, no. 2, pp. 285–296, Feb. 2024, doi: 10.30812/matrik.v23i2.3352.
[25] G. Sugendran and S. Sujatha, “Earlier identification of heart disease using enhanced genetic algorithm and fuzzy weight based support vector machine algorithm,” *Meas. Sensors*, vol. 28, p. 100814, Aug. 2023, doi: 10.1016/j.measen.2023.100814.
[26] P. Jeemon *et al.*, “Efficacy of a family-based cardiovascular risk reduction intervention in individuals with a family history of premature coronary heart disease in India (PROLIFIC): an open-label, single-centre, cluster randomised controlled trial,” *Lancet Glob. Heal.*, vol. 9, no. 10, pp. e1442–e1450, Oct. 2021, doi: 10.1016/S2214-109X(21)00319-3.