Jurnal Varian
Vol. 8, No. 3, October 2025, pp 307–318
e-ISSN: 2581-2017

Ensemble Quick Robust Clustering Using Links for Clustering Hypertension Patients at a Health Center

Neli Niftayana, Mohammad Fajri, Nurul Fiskia Gamayanti
Universitas Tadulako, Palu, Indonesia

Article Info

Article history:

Received : 02-19-2025
Revised : 03-20-2025
Accepted : 06-01-2025

Keywords:

Agglomerative Nesting;
Clustering;
Ensemble;
Hypertension;
Quick Robust Clustering Using Links.

ABSTRACT

Hypertension is a chronic disease with a high risk of cardiovascular complications and requires treatment according to patient characteristics. At the health center, the number of hypertensive patients is 6953, the highest recorded. Therefore, this study aims to classify and determine the characteristics of hypertensive patients at a health center. The method used in this study is Ensemble Quick Robust Clustering Using Links. This method combines the clustering results of Quick Robust Clustering Using Links and Agglomerative Nesting. Where this method is more efficient in clustering. The results of this study show the number of clusters in the Quick Robust Clustering Using Links method is 3, Agglomerative Nesting is 3 and in the Quick Robust Clustering Using Links Ensemble produces 9 clusters with the following distribution: Cluster 1 shows low hypertension, cluster 2 shows high hypertension, cluster 3 to cluster 6 shows high hypertension, cluster 7 shows moderate hypertension, cluster 8 shows high hypertension and cluster 9 shows moderate hypertension. Thus, grouping patients based on a combination of numerical and categorical variables can provide more detailed information about the severity of hypertension.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v8i3.5151

Corresponding Author:

Neli Niftayana,
Department of Statistics, Universitas Tadulako, Palu, Indonesia,
Email: neliniftayana@gmail.com

How to Cite:
Niftayana, N., Fajri, M., & Gamayanti, N. F. (2025). Ensemble Quick Robust Clustering Using Links for Clustering Hypertension Patients at a Health Center. *Jurnal Varian*, *8*(3), 307–318.

A. INTRODUCTION

Hypertension represents a major public health issue and is recognized as one of the leading contributors to premature mortality worldwide. A diagnosis is established when blood pressure measurements taken on two separate occasions show systolic values of $\geq$140 mmHg and/or diastolic values of $\geq$90 mmHg. Effective management and proper control of hypertension can improve overall quality of life and significantly lower the likelihood of complications such as coronary artery disease, heart failure, stroke, and chronic kidney diseas (Nawi et al., 2021). Hypertension is acknowledged as one of the most critical risk factors contributing to both morbidity and mortality at the global level, accounting for an estimated nine million deaths annually. The emergence and rapid development of modern technologies have begun to reshape approaches in the diagnosis, monitoring, and management of this condition. In particular, the use of smartphones, wearable health devices, and telemonitoring systems is showing significant potential to improve how blood pressure is measured, tracked over time, and managed in daily clinical practice (Kitt et al., 2019).

In Indonesia, the prevalence of hypertension continues to increase. According to the Basic Health Research (Widyawati, 2018), the prevalence of hypertension in Indonesia reached 34.1%. Regions outside Java Island, including Central Sulawesi, often face challenges in the equitable distribution of healthcare services. This study highlights the importance of regional analysis to identify areas with urgent healthcare needs, such as Donggala, which has a high prevalence of hypertension but limited healthcare services. According to data from the Central Sulawesi Provincial Health Office Profile, there are 384,072 people with hypertension in this province, or about 2.33% of the population. The highest percentage of hypertension patients based on the largest disease estimate occurred in 2020 in the Donggala District, with a value of 7.11%. From this data, it is known that 65,398 people are suffering from hypertension, but only 4,650 people receive hypertension care services (Dinas Kesehatan Provinsi Sulawesi Tengah, 2021).

According to data from the Health Profile of the Sabang Community Health Center in Dampelas District, Donggala Regency, the number of people aged $\geq$15 years with hypertension at the Ita Seseibi Sabang Community Health Center was 6,953 (6.9%), which was the highest number compared to other non-communicable diseases. The highest hypertension prevalence rate in 2021 was recorded in Karya Mukti Village, with a prevalence rate of 15.5%. Based on this data, it is estimated that there are 1,057 hypertensive patients aged 15 years and older in Karya Mukti Village, but only 57 of them received hypertension-related services. Conversely, Kambayang Village has the lowest hypertension rate, with an estimated 232 hypertensive patients, of whom 16 received services. The high number of hypertension patients is closely related to the lifestyle of the community, which tends to involve minimal physical activity, excessive consumption of high-sodium foods, excessive caffeine intake, smoking, being overweight or obese, dyslipidemia, and experiencing stress.

The purpose of this method is to identify the severity of a patient's hypertension, utilizing information obtained through various classification models. The method that can be used with mixed numerical and categorical data is the Ensemble method. Numerical data is grouped using the Hierarchical Agglomerative Nesting (AGNES) algorithm, while categorical data is grouped using Quick Robust Clustering Using Links (QROCK). After that, the results of these two groupings are combined (Ensemble) and processed with a clustering algorithm for categorical data (Ensemble QROCK) to obtain the final result. This algorithm is known as algCEBMDC (Shofari, 2024). The algCEBMDC method is a clustering analysis using the Cluster Ensemble approach, which divides the initial mixed-type data into two distinct sub-datasets: pure numerical data and pure categorical data. Each sub-dataset is then analyzed and clustered, and the clustering results are combined to produce the final cluster. One of the Ensemble methods that is often used is Robust Clustering Using Links (ROCK).

In a study by Sari & Saputro (2021), the QROCK algorithm was evaluated in terms of accuracy and efficiency. The findings indicate that it is superior in terms of efficiency and accuracy due to its ability to detect and handle outliers in categorical data well, as shown in a study by Hermanto et al. (2024) analyzed the grouping of 100 hypertensive patient data at the Bungah Community Health Center, using a Web-Based K-Means Clustering Algorithm. The results of the study show that the 100-hypertension data were grouped into 4 clusters: cluster 1, Isolated Systolic Hypertension; cluster 2, Grade 1 (mild hypertension); cluster 3, Grade 2 (moderate hypertension); and cluster 4, Grade 3 (severe hypertension). The hypertension data used consisted of two attributes, namely systolic and diastolic.

The gap between this study and previous ones is that most prior studies used only a single clustering method and were limited to systolic and diastolic blood pressure variables, resulting in simple clustering results. The difference between this study and previous ones is that it uses the QROCK ensemble approach combined with Agglomerative Nesting and includes additional variables such as age, body mass index, cholesterol, genetic history, smoking habits, salt consumption, fat consumption, and physical activity. Thus, the clustering results obtained are more comprehensive in describing the characteristics of hypertensive patients. The purpose of this study is to classify hypertensive patients based on relevant numerical and categorical variables, thereby providing more in-depth information for targeted hypertension prevention and treatment efforts.

The contribution of this research is to support hypertension control efforts, which require an appropriate data analysis approach. The use of clustering methods is one promising solution for identifying the severity of hypertension. Clustering methods based on mixed data, such as algCEBMDC, which combines the AGNES algorithm for numerical data and QROCK for categorical data, are considered capable of providing more accurate and efficient results. As such, this method has the potential to assist healthcare professionals in designing more targeted interventions based on patient characteristics.

B. RESEARCH METHOD

The data used in this study are secondary data sourced from the medical record data of a Health Center. This study used 100 data on hypertensive patients at a Health Center in the period 2024. The 10 independent variables used in this study are age (X_1) the incidence of hypertension increases with age (Zubiyo et al., 2025), body mass index (X_2) a measurement based on a person's

height and weight (Khanna et al., 2022), systolic (X_3) & diastolic (X_4) is a condition in which blood pressure is ≥140 mmHg (systolic pressure) and or ≥90 mmHg (diastolic pressure) (Nurlaela et al., 2025), cholesterol (X_5) levels below or equal to 199 mg/dL are considered normal or good, 200 to 239 mg/dL is moderate, while values above 240 mg/dL are considered high and can be dangerous (Niran, 2024), genetic (X_6) a family history of hypertension in parents plays an important role in predicting the hypertensive phenotype in their offspring (Zhao et al., 2021), smoking (X_7) is one of the biggest risk factors that can cause disease and death, one of which is hypertension (Jareebi, 2024), more salt (X_8) can cause an increase in blood pressure, which contributes to cardiovascular morbidity and mortality (Bailey & Dhaun, 2024), more fat (X_9) can cause excessive and abnormal cholesterol levels in the blood (Jiang et al., 2016), physical activity (X_{10}) consists of any movement that involves skeletal muscles and requires energy expenditure (Singh et al., 2020). The data was then processed with the ensemble quick robust clustering method, employing links r, through the following steps, as shown in the flowchart in Figure 1.

Data analysis in this study was conducted using RStudio software. The following are the stages of analysis carried out in this study:

1. Input data used as research variables.
2. Cluster Analysis
   a. Measuring the distance of numeric type data using Euclidean distance.

      Min-max normalization is a normalization method that uses a linear strategy to transform data from one range of values to a new range of values. This process results in a balance in the comparison value between the data before and after normalization (Prasad & T, 2024). The formula for min-max normalization is defined as shown in Equation (1):

$$normalized\,(Z) = \frac{x - min(x)}{max(x) - min(x)} \tag{1}$$

      Where Z represents the normalization result, x is the value (original), $\min(x)$ is the minimum value for variable x, and $\max(x)$ is the maximum value for variable x. This method transforms the original data into a range between 0 and 1 by scaling based on the minimum and maximum values of the variable.

      Euclidean distance, which calculates the distance between observations (Faizah et al., 2020) as shown in Equation (2):

$$d\,(i,j) = \sqrt{\sum_{k=1}^{n} (X_i - X_j)^2} \tag{2}$$

      Where $d(i,j)$ represents the distance between the i object and the j object, X_i is the i object vector, and X_j is the j object vector. The commonly used measurement of numerical variable distance dissimilarity is Euclidean distance.

   b. Perform distance measurement of categorical type data using a weighted similarity measure.

      The weighted similarity measure between two objects is shown in Equation (3) (Shofari, 2024):

$$sim\,(X_i X_j) = \frac{|X_i \cap X_j|}{|X_i \cap X_j| + 2\sum_{K \notin X_i \cap X_j} \frac{1}{|D_k|}} \tag{3}$$

      Where $|X_i \cap X_j|$ represents the number of categories in common between X_i and X_j, and $|D_k|$ is the difference in category level between X_i and X_j. Which $sim\,(X_i X_j)$ denotes the similarity between the object pairs X_i and X_j.

3. Numeric Type Data Grouping
   a. Perform clustering using the agglomerative hierarchy method (single linkage, complete linkage, and average linkage) (Abushilah & Abbas, 2023):
      1) Single linkage is based on the smallest or closest distance. The distance is measured as shown in Equation (4):

$$d_{(ij)k} = \min(d_{ik,} d_{jk}) \tag{4}$$

         Where d_{ik} represents the closest distance from cluster i and k, and d_{jk} is the distance from cluster j and k.

      2) Complete linkage is based on the largest distance or the farthest distance. The distance measured as shown in Equation (5):

$$d_{(ij)k} = \max(d_{ik}, d_{jk}) \tag{5}$$

Where d_{ik} represents the farthest distance from the cluster i and k, and d_{jk} The distance is the farthest from the cluster j and k.

3) Average linkage is based on averaging all distances between objects. The distance measured as shown in Equation (6):

$$d_{(ij)k} = \frac{\sum_{i=1}^{n} \sum_{k=1}^{n} d_{ik}}{n_{(ij)} n_k} \tag{6}$$

Where d_{ik} represents the distance of object i in cluster (uv) with object j in cluster k, $n_{(ij)}$ is the number of objects in cluster (ij) n_k number of objects in cluster k, and n_k is the number of objects in cluster k.

b. Determining the better model by validating the cluster based on the cophenetic correlation coefficient obtained.

The average value of each cluster on each variable can provide insight into the characteristics of the cluster (Kumar & Toshniwal, 2016). In addition, the kophenetic correlation is used to evaluate how well or poorly a clustering object is placed in a cluster, which helps in determining the optimal cluster result. Based on the kophenetic correlation coefficient, the following equation is shown in Equation (7):

$$r_{coph} = \frac{\sum_{i<j}(d_{ij} - d)(d_{cij} - d_c)}{\sqrt{[\sum Z_{i<j}(d_{ij} - d)^2] \ [\sum_{i<i}(d_{cij} - d_c)^2]}} \tag{7}$$

Where r_{Coph} represents the cophenetic correlation coefficient, d_{ij} is the original distance (Euclidean distance) between objects i and j, d is the average d_{ij}, dc_{ij} is the cophenetic distance of objects i and j, and dc is the average dc_{ij}. The value of this coefficient ranges from 0 to 1, with values closer to 1 indicating a better fit.

4. Categorical Type Data Grouping

a. Perform clustering using the QROCK method with a threshold value (θ) between 0 and 1.

b. Calculating the ratio of $S\prime_w$ and $S\prime_B$ to determine the optimum number of clusters based on categorical data.

The performance of clustering results for categorical scale variables is known from ANOVA (analysis of variance) using equivalent contingency tables. If there are n observations, with n_k is the number of observations in the kth category where $k = 1, 2, 3, \ldots, K$ and $n = \sum_{k=1}^{K} n_k$. Furthermore n_{kc} is the number of observations with the kth category and the cth cluster, where $c = 1, 2, 3, \ldots, C$ with C is the number of clusters formed, so that $n_c = \sum_{k=1}^{K} n_{kc}$ is the number of observations in the c-th cluster and $n_k = \sum_{c=1}^{C} n_{kc}$ is the number of observations in the k-th category. So the total number of observations can be written as $n = \sum_{c=1}^{C} n_c = \sum_{k=1}^{K} n_k = \sum_{k=1}^{K} \sum_{c=1}^{C} n_{kc}$. So the equations are (Alvionita, 2017):

The Total Squared Sum (SST) of categorical data variables is formulated in the following Equation (8):

$$SST = \frac{n}{2} - \frac{1}{2n} \sum_{k=1}^{K} n_k^2 \tag{8}$$

The Total Sum of Squares Within Cluster (SSW) is formulated in the following Equation (9):

$$SSW = \sum_{c=1}^{C} \left(\frac{n_c}{2} - \frac{1}{2n_c} \sum_{k=1}^{K} n_{kc}^2 \right) = \frac{n}{2} - \frac{1}{2n} \sum_{c=1}^{C} \frac{1}{n_c} \sum_{k=1}^{K} n_{kc}^2 \tag{9}$$

The total sum of squares between clusters (SSB) is formulated in the following Equation (10):

$$SSB = \frac{1}{2} \left(\sum_{c=1}^{C} \frac{1}{n_c} \sum_{k=1}^{K} n_{kc}^2 \right) - \frac{1}{2n} \sum_{k=1}^{K} n_k^2 \tag{10}$$

Whitin's Mean of Squares (MSW) is formulated in the following Equation (11):

DOI: https://doi.org/10.30812/varian.v8i3.5151

$$MST = \frac{SST}{(n-1)} \tag{11}$$

Mean of Squares Within (MSW) is formulated in the following Equation (12):

$$MSW = \frac{SSW}{(n-C)} \tag{12}$$

Mean of Squares Between (MSB) is formulated in the following Equation (13):

$$MSB = \frac{SSB}{(C-1)} \tag{13}$$

The standard deviation within clusters (S_W) and standard deviation between clusters (S_B) of categorical data are formulated in the following Equations (14)-(15):

$$S\prime_W = [MSW]^{\frac{1}{2}} \tag{14}$$

$$S\prime_B = [MSB]^{\frac{1}{2}} \tag{15}$$

The clustering performance of categorical data is based on the ratio between the standard deviation within clusters (S_W) and the standard deviation between clusters (S_B). If the ratio is smaller, the clustering performance of categorical data improves, resulting in maximum homogeneity within clusters and maximum heterogeneity between clusters.

5. Final Cluster Grouping QROCK Ensemble Method

The clustering analysis on mixed numerical and categorical scale data was initially divided into two subdata sets: pure numerical data and pure categorical data. Next, data clustering was performed separately for each type of data. The results of each clustering are then combined using the QROCK Ensemble cluster method to obtain the final clusters. The steps in analyzing mixed data using the Ensemble cluster method, called the CEBMDC algorithm, have the following stages (Ratnasari & Dani, 2023):

1) Divide the mixed data into two data parts: purely numeric data and purely categorical data.
2) Perform clustering with numeric data clustering algorithms for numeric variables, and perform clustering with categorical data clustering algorithms for categorical variables.
3) Combining the clustering results (outputs) of numerical variables and categorical variables. This combining is called the Ensemble process.
4) Perform Ensemble clustering using categorical data clustering algorithms to obtain final clusters.

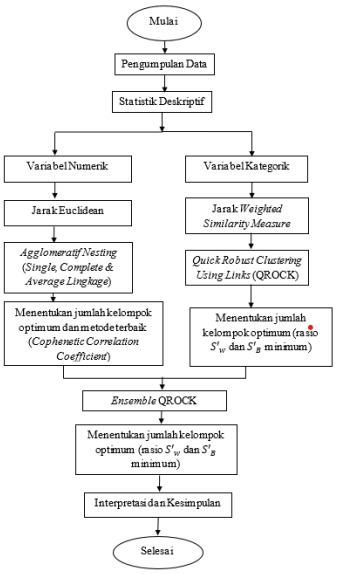

Figure 1. Conceptual Framework of the Research

C. RESULT AND DISCUSSION

1. Cluster Analysis

The results of cluster analysis are influenced by the variables observed, the size scale used, and the clustering method used. Clustering is done based on similarity or dissimilarity between objects of observation; the size of similarity and dissimilarity is generally measured based on distance.

a. Numerical Distance Measurement

Before performing numerical distance measurements, normalization is first performed to ensure that each feature has an equal contribution in the analysis or modeling. Normalization is very important, especially for those who use distance. The results of data normalization can be seen in Table 1 below.

Table 1. Normalization Result Data

Patient	X_1	X_2	X_3	X_4	X_5
P1	0.865	-0.175	2.423	0.788	1.308
P2	1.457	0.461	7.743	3.482	5.549
P3	-1.069	-1.391	0.128	-0.849	0.372
P4	0.218	-0.383	1.397	0.564	1.821
⋮	⋮	⋮	⋮	⋮	⋮
P99	0.089	-0.409	1.474	0.564	2.667
P100	-0.121	-0.296	0.495	0.101	0.187

In the Agnes clustering process, the first step is to randomly initiate the cluster center, followed by calculating the distance from each data point to the predetermined cluster center. This calculation can be done using the Euclidean distance matrix. Table 2 presents the Euclidean distance matrix.

Table 2. Euclidean Distance Matrix

Patient	P1	P2	P3	⋯	P99	P100
P1	0	7.369	3.748	⋯	1.858	2.537
P2	7.369	0	10.648	⋯	7.665	9.787
P3	3.748	10.648	0	⋯	3.374	1.781
⋮	⋮	⋮	⋮	⋱	⋮	⋮
P99	1.858	7.665	3.374	⋯	0	2.717
P100	2.537	9.787	1.781	⋯	2.717	0

Based on Table 2, the distance from each data point to the cluster center, which was randomly determined at the beginning, can be calculated. These results can be used to determine cluster members, where each data point with the smallest distance to a particular cluster center will be grouped into that cluster.

b. Categorical Distance Measurement

Initiate the use of a threshold value as a limit for determining neighbors, with values such as 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. If the distance between observations exceeds the threshold value, it indicates that the observations are considered neighbors, leading to the merging of object components x and y if one object is a neighbor of the other. The Weighted Similarity Measure Distance results are presented in Table 3.

Table 3. Euclidean Distance Matrix

Patient	P1	P2	P3	⋯	P100
P1	1	0.6	0.6	⋯	1
P2	0.6	1	1	⋯	0.6
P3	0.6	1	1	⋯	0.6
⋮	⋮	⋮	⋮	⋱	⋮
P99	1	0.6	0.6	⋯	1
P100	1	0.6	0.6	⋯	1

DOI: https://doi.org/10.30812/varian.v8i3.5151

Based on Table 3, the distance from each data point to the cluster center is determined by the distance between observations exceeding the threshold value, indicating that these observations are considered neighbors.

2. Numerical Data Grouping

The ideal number of groups for the three linkage methods —single linkage, complete linkage, and average linkage —is determined by the maximum Cophenetic Correlation Coefficient Value for each method. The results of the Cophenetic Correlation Coefficient are presented in Table 4.

Table 4. Cophenetic Correlation Coefficient r_{coph}

Methods	Cophenetic Correlation Coefficient
Single	0.9155304
Complete	0.9208933
Average	0.9271516

Based on Table 4 above, it is known that the highest cophenetic correlation coefficient is in the Average linkage clustering method. The cophenetic correlation coefficient measures how well the dendrogram (the result of hierarchical clustering) represents the original numerical data of the distances between hypertension patients at a health center.

3. Categorical Data Grouping

Categorical data is clustered using the QROCK method using a threshold value as a limit for determining neighbors, with values of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. is selected from the ten predetermined threshold values. The results of the SW and SB ratios are presented in Table 5.

Table 5. Results of S_W and S_B Ratio QROCK Grouping

Threshold	Number of Clusters	S_W	S_B	Ratio
0.1	1	15.846	-	-
0.2	1	15.846	-	-
0.3	1	15.846	-	-
0.4	1	15.846	-	-
0.5	1	15.846	-	-
0.6	1	15.846	-	-
0.7	3	15.986	59.552	0.2684
0.8	3	15.986	59.552	0.2684
0.9	7	16.313	42.523	0.3836

Based on Table 5, the ratio value between SW and the smallest SB is at a threshold of 0.7 and 0.8, namely 0,2684. Clustering hypertensive patients at a health center, using categorical variables, resulted in a total of 3 clusters. This relatively small ratio indicates that the variation between clusters is greater than the variation within clusters, leading to the conclusion that the clustering results show a better and clearer structure.

4. Mixed Data Grouping

The results of grouping numerical and categorical data that have been obtained are then grouped by viewing the results as new categorical data (Table 6):

Table 6. Mixed Data of Numeric and Categorical Grouping Results

Patient	Cluster Categoric	Cluster Numeric
P1	1	1
P2	2	2
P3	2	1
P4	2	1
P5	2	1
P6	1	1
P7	1	3
P8	2	1

DOI: https://doi.org/10.30812/varian.v8i3.5151

Patient	Cluster Categoric	Cluster Numeric
P9	2	1
⋮	⋮	⋮
P94	2	1
P95	2	1
P96	2	1
P97	2	3
P98	1	1
P99	1	1
P100	1	1

The combined result data is clustered using the QROCK method. Initiate thresholds as neighbor boundary thresholds, with values of 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, and 0.9. The smallest value of the ratio of S_W and S_B is selected from the ten predefined threshold values.

Table 7. Results of S_W and S_B Ratio of QROCK Ensemble Grouping

Treshold	Number of Clusters	S_W	S_B	Ratio
0.1	1	0.8910	-	-
0.2	1	0.8910	-	-
0.3	1	0.8910	-	-
0.4	1	0.8910	-	-
0.5	1	0.8910	-	-
0.6	1	0.8910	-	-
0.7	9	10.430	24.339	0.4285
0.8	9	10.430	24.339	0.4285
0.9	9	10.430	24.339	0.4285

Based on Table 7, it can be seen that at threshold values of 0.1 to 0.6, the QROCK ensemble method forms one cluster. This shows that at low similarity levels, the algorithm has not been able to distinguish the data structure significantly, so all objects are considered to be in the same group. In this condition, the value of variance within clusters (S_W), variance between clusters (S_B), and the ratio of the two cannot be calculated because there is no differentiator between clusters. Changes began to appear at thresholds of 0.7 and 0.9, where the number of clusters increased to nine. The same SW and SB values at the three thresholds, 1.0430 and 2.4339, respectively, resulted in a ratio of 0.4285. This relatively small ratio indicates that the variation between clusters is greater than the variation within clusters, suggesting that the clustering results are beginning to show a clearer and more structured pattern. These results are consistent with the findings of Sari & Saputro (2021) study, which showed that QROCK is effective in handling categorical data with a high level of accuracy.

5. Patient Characteristics of QROCK Ensemble Grouping Results

The Quick Robust Clustering Using Links ensemble clustering results that have been obtained contain the division of hypertension patients based on the QROCK ensemble clustering results:

Table 8. Patient Distribution Based on Clustering Results

Cluster	Patient
Cluster 1	P1, P6, P24, P25, P26, P28, P34, P35, P36, P38, P39, P40, P41, P43, P65, P66, P69, P70, P71, P73, P74, P79, P80, P81, P89, P90, P91, P93, P98, P99 dan P100.
Cluster 2	P2.
Cluster 3	P3, P4, P5, P8, P9, P10, P11, P13, P14, P15, P16, P18, P19, P20, P21, P23, P46, P48, P49, P50, P51, P53, P54, P55, P60, P61, P63, P64, P68, P83, P84, P85, P86, P88, P94, P95 DAN P96.
Cluster 4	P7, P27, P37, P42, P62, P67, P72 dan P92.
Cluster 5	P12, P17, P22, P47, P52, P87 dan P97.
Cluster 6	P29, P30, P31, P33, P44, P45, P56, P58, P59, P75, P76 dan P78.

Cluster	Patient
Cluster 7	P32 dan P57.
Cluster 8	P77.
Cluster 9	P82.

Based on Table 8, the findings of this study indicate that the QROCK ensemble method is capable of generating nine clusters of hypertensive patients with different characteristics. Cluster 1 showed a relatively stable physiological condition, with blood pressure and cholesterol within reasonable limits and body mass index close to the healthy category. The lifestyle of this group is relatively good, with no family history of hypertension, no smoking, and physically active; this group is classified as low risk. Cluster 2 had very high blood pressure and mild obesity. Cholesterol was also high. The lifestyle of this group has no history of hypertension or smoking, and poor diet and low physical activity dominate this group. With a heavy physiological burden and an unhealthy lifestyle, this group is categorized as high risk. Cluster 3 shows moderately high blood pressure and cholesterol, despite a moderate body mass index. The lifestyle of this group tends to be unhealthy, characterized by excessive consumption of salt and fat and a lack of physical activity. However, without a history of hypertension, this group is in the high-risk category.

Furthermore, Cluster 4 shows high blood pressure and cholesterol conditions, with a body mass index in the moderate category. Family history of hypertension is not very prominent, but unhealthy diet and low activity are the main characteristics; this group has the potential to experience complications of hypertension. Cluster 5 has a combination of high blood pressure, cholesterol, and body mass index. An unhealthy lifestyle is dominant, particularly in the consumption of foods high in salt and fat and a lack of physical activity; this group is at high risk. Cluster 6 shows high blood pressure and cholesterol with a tendency to be overweight. A less health-supportive lifestyle is evident in the unbalanced diet and low physical activity, although there are no genetic risk factors, this group is at high risk. Cluster 7 had relatively lower blood pressure than the other groups and a healthy body mass index. However, there were strong genetic risk factors and high consumption habits of salt and fat. The physical condition seemed better, and this group was categorized as moderate risk. Cluster 8 had a high mean score, with very high blood pressure and cholesterol, and a tendency to be overweight. All aspects of lifestyle were classified as very poor, from family history of hypertension, smoking, excessive consumption of salt and fat, to lack of physical activity. This group was categorized as high. Cluster 9 shows high blood pressure and low body mass index. There is no family history of hypertension or smoking, but diet remains a major risk factor; this group is classified as moderate risk. These findings are consistent with or supported by the research of Hermanto et al. (2024), who also found that the clustering method was able to distinguish hypertension levels into several categories.

D. CONCLUSION AND SUGGESTION

In 2024, clustering hypertension patients at a center based on numeric (age, body mass index, systolic and diastolic blood pressure, cholesterol) and categorical (hereditary history, smoking, salt and fat consumption, physical activity) variables using the Quick Robust Clustering Using Links ensemble resulted in 9 clusters. Cluster 1 has 31 patients, cluster 2 has 1 patient, cluster 3 has 37 patients, cluster 4 has 8 patients, cluster 5 has 7 patients, cluster 6 has 12 patients, cluster 7 has 2 patients, cluster 8 has 1 patient and cluster 9 has 1 patient.

In determining the characteristics of each cluster based on their profiling values, it can be seen that cluster 1 indicates patients generally have a low risk of hypertension. Cluster 2 indicates that patients in this cluster generally have a high risk of hypertension. Cluster 3 indicates a hypertension group with a high risk category. Cluster 4 indicates the potential to experience complications of hypertension. Cluster 5 and Cluster 6 have a high-risk hypertension group. Cluster 7 has a moderate risk of hypertension. Cluster 8 has a hypertension group with a high-risk category. Cluster 9 shows hypertension that is classified as moderate.

ACKNOWLEDGEMENT

The author would like to thank all parties that gave contribution to this research. Hopefully this research could make a significant impact to society and science.

DECLARATIONS

AUTHOR CONTIBUTION

In this study, first author contributed as a writer, conducted the data analysis and made revisions to the reviewer suggestions. The

second author provides research topic ideas and undertake data interpretation. The third author prepared the initial draft and assisted in collecting data.

FUNDING STATEMENT

This research is self-funded by the authors.

COMPETING INTEREST

The authors firmly assert that there is no conflict of interest in this article.

REFERENCES

Abushilah, S. F., & Abbas, R. H. (2023). Performance Evaluation of Some Clustering Algorithms under Different Validity Indices. *Mathematical Modelling of Engineering Problems*, *10*(4), 1271–1280. https://doi.org/10.18280/mmep.100420

Alvionita, A. (2017, March 18). *Metode ensembel rock dan swfm untuk pengelompokan data campuran numerik dan kategorik pada kasus aksesi jeruk* [Thesis]. Institut Teknologi Sepuluh Nopember. https://repository.its.ac.id/2440/

Bailey, M. A., & Dhaun, N. (2024). Salt Sensitivity: Causes, Consequences, and Recent Advances. *Hypertension*, *81*(3), 476–489. https://doi.org/10.1161/HYPERTENSIONAHA.123.17959

Dinas Kesehatan Provinsi Sulawesi Tengah. (2021). *Profil Kesehatan Sulawesi Tengah 2021*. https://dinkes.sultengprov.go.id/wp-content/uploads/2022/05/PROFIL-DINAS-KESEHATAN-2021.pdf

Faizah, N., Surohman, Fabrianto, L., Hendra, & Prasetyo, R. (2020). Unbalanced Data Clustering with K-Means and Euclidean Distance Algorithm Approach Case Study Population and Refugee Data. *Journal of Physics: Conference Series*, *1477*(2), 022005. https://doi.org/10.1088/1742-6596/1477/2/022005

Hermanto, H., Hendi, A., & Zuhriyah, A. (2024). Implementation of the Web-Based K-Means Clustering Algorithm on Hypertension Levels in the Elderly at the Bungah District Health Center. *Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics*, *6*(2), 65–77. https://doi.org/10.35882/h6596074

Jareebi, M. (2024). The Association Between Smoking Behavior and the Risk of Hypertension: Review of the Observational and Genetic Evidence. *Journal of Multidisciplinary Healthcare*, *17*, 3265–3281. https://doi.org/10.2147/JMDH.S470589

Jiang, S.-Z., Lu, W., Zong, X.-F., Ruan, H.-Y., & Liu, Y. (2016). Obesity and hypertension. *Experimental and Therapeutic Medicine*, *12*(4), 2395–2399. https://doi.org/10.3892/etm.2016.3667

Khanna, D., Peltzer, C., Kahar, P., Parmar, M. S., Khanna, D., Peltzer, C., Kahar, P., & Parmar, M. S. (2022). Body Mass Index (BMI): A Screening Tool Analysis. *Cureus*, *14*(2). https://doi.org/10.7759/cureus.22119

Kitt, J., Fox, R., Tucker, K. L., & McManus, R. J. (2019). New Approaches in Hypertension Management: A Review of Current and Developing Technologies and Their Potential Impact on Hypertension Care. *Current Hypertension Reports*, *21*(6), 44. https://doi.org/10.1007/s11906-019-0949-4

Kumar, S., & Toshniwal, D. (2016). Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). *Journal of Big Data*, *3*(1), 13. https://doi.org/10.1186/s40537-016-0046-3

Nawi, A. M., Mohammad, Z., Jetly, K., Razak, M. A. A., Ramli, N. S., Wan Ibadullah, W. A. H., & Ahmad, N. (2021). The Prevalence and Risk Factors of Hypertension among the Urban Population in Southeast Asian Countries: A Systematic Review and Meta-Analysis (M. Salvetti, Ed.). *International Journal of Hypertension*, *2021*, 1–14. https://doi.org/10.1155/2021/6657003

Niran, A. (2024). A Review of Cholesterol's Dual Impact: Human Health and Environmental Consequences. *Badeggi Journal of Agricultural Research and Environment*, *6*(2), 144–155. https://doi.org/10.35849/BJARE202402/189/013

Nurlaela, G., Taobah Ramdani, H., & Wahyudin. (2025). Nursing Care Analysis of Hypertension Through Hypertension Management with Swedish Massage Therapy. *Nursing Case Insight Journal*, *3*(1), 30–33. https://doi.org/10.63166/zmz8bn56

Prasad, M., & T, S. (2024, November 7). *Clustering Accuracy Improvement Using Modified Min-Max Normalization Technique*. https://doi.org/10.20944/preprints202411.0486.v1

Ratnasari, V., & Dani, A. T. R. (2023). Mapping the Provincial Food Security Conditions in Indonesia Using Cluster Ensemble-Based Mixed Data Clustering-Robust Clustering with Links (CEBMDC-ROCK). *International Journal on Advanced Science, Engineering and Information Technology*, *13*(2), 611–617. https://doi.org/10.18517/ijaseit.13.2.16457

Sari, I. A., & Saputro, D. R. S. (2021). Algoritme Quick RObust Clustering using linKs (QROCK) untuk Clustering Data Kategorik. *PRISMA, Prosiding Seminar Nasional Matematika*, *4*, 640–644. https://journal.unnes.ac.id/journals

Shofari, M. R. (2024). Quick Robust Clustering Using Links (QROCK) untuk Pengelompokan Desa Kabupaten Banjar. *RAGAM: Journal of Statistics & Its Application*, *3*(1), 97. https://doi.org/10.20527/ragam.v3i1.12805

Singh, R., Pattisapu, A., & Emery, M. S. (2020). US Physical Activity Guidelines: Current state, impact and future directions. *Trends in Cardiovascular Medicine*, *30*(7), 407–412. https://doi.org/10.1016/j.tcm.2019.10.002

Widyawati, W. (2018, November 2). *Potret kesehatan indonesia dari riskesdas 2018*. Kemenkes. https://kemkes.go.id/id/potret-sehat-indonesia-riskesdas-2018

Zhao, W., Mo, L., & Pang, Y. (2021). Hypertension in adolescents: The role of obesity and family history. *The Journal of Clinical Hypertension*, *23*(12), 2065–2070. https://doi.org/10.1111/jch.14381

Zubiyo, A., Rohani, T., & Rustandi, H. (2025). Factors Associated With The Incidence Of Hypertension In Young Adults At Pasar Ikan Community Health Centre, Bengkulu City, 2024. *Multidisciplinary Journals*, *2*(2), 57–66. https://doi.org/10.37676/mj.v2i2.696

[This page intentionally left blank.]

Vol. 8, No. 3, October 2025, pp 307–318
DOI: https://doi.org/10.30812/varian.v8i3.5151