Jurnal Aplikasi Sains Data
Journal homepage: https://jasid.upnjatim.ac.id/index.php/jasid

Comparative Analysis of Hierarchical Clustering and K-Medoids for Clustering Cases of Childhood Respiratory Diseases in Lamongan Regency

Adelia Yuandhika[1], Nezalfa Sabrina[2], Cahya Eka Melati[3], Dwi Arman Prasetya[4], and Prismahardi Aji Riyantoko[5,*]

[1,2,3,4]UPN "Veteran" Jawa Timur, [5]Okayama University

[1]22083010066@student.upnjatim.ac.id, [2]22083010067@student.upnjatim.ac.id, [3]22083010090@student.upnjatim.ac.id, [4]arman.prasetya.sada@upnjatim.ac.id, [5,*]pnai2m3s@s.okayama-u.ac.jp

ABSTRACT

Abstract— Respiratory diseases affecting children remain a significant health issue in Indonesia, including in Lamongan Regency. The region faces challenges related to pediatric respiratory illnesses, particularly Childhood Tuberculosis, Pneumonia in toddlers, and Cough in toddlers, which impact children's quality of life and development. Therefore, understanding the spatial distribution and correlation patterns among these diseases is essential to support more targeted health intervention planning. This study analyzes the distribution patterns of pediatric respiratory diseases in Lamongan Regency and clusters regions based on similarities in the number of cases using an unsupervised learning approach. The method employed is Hierarchical Clustering with four distance calculation techniques: single, complete, average, and ward linkage and K-Medoids with two distance calculation techniques: euclidean and manhattan distance. The data, sourced from the Lamongan District Health Office, include four numerical variables related to respiratory diseases, aggregated by sub-districts. Data normalization was carried out using standardization, and cluster quality was evaluated using three internal metrics: Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). The analysis results indicate that the optimal number of clusters is three. Among all methods tested, the Hierarchical Clustering with ward linkage method yielded the best performance, with a Silhouette Score of 0.5447, a DBI of 0.5884, and a CHI of 20.3018. These results demonstrate that the ward linkage method is the most effective in clustering regions based on the characteristics of pediatric respiratory disease cases and can be used for mapping priority health intervention areas in Lamongan Regency.

Keywords: Pediatric Respiratory Diseases, Hierarchical Clustering, Ward Linkage, K-Medoids.

I. INTRODUCTION

Diseases affecting the human respiratory system are one of the health problems that are still a serious concern in Indonesia. The respiratory system plays a vital role in sustaining life, so disruptions to this system can have a serious impact, especially on vulnerable groups such as children. Respiratory diseases such as pneumonia and pediatric tuberculosis are listed as the leading causes of morbidity and mortality in children under five. Based on data from the Ministry of Health of the Republic of Indonesia, pneumonia accounts for approximately 14.5% of infant deaths and 5% of under-five deaths each year [1].

Lamongan district faces significant challenges related to childhood respiratory diseases, especially cases of childhood tuberculosis, pneumonia among children under five, non-pneumonia cough, and cough in children under five. These cases have a high incidence and impact on children's development

* *Corresponding author.*
E-mail address: pnai2m3s@s.okayama-u.ac.j
10.33005/jasid.v2i1.37

and quality of life. However, until now there has been no study that specifically analyzes the spatial distribution or patterns of interrelationships between types of respiratory diseases in children in the region. Therefore, a data-driven approach is needed to identify disease distribution patterns more comprehensively to support the formulation of targeted health interventions.

One approach that can be used to reveal patterns of similarity between regions based on disease data is clustering analysis. Clustering is an explorative method in unsupervised learning that is used to group data based on similar characteristics. In this study, two clustering algorithms are used, namely Hierarchical Clustering, which constructs a hierarchical structure based on the closeness between data, and K-Medoids, which uses medoids as cluster centers and is known to be more resistant to outliers [2].

This study aims to analyze the clustering of respiratory disease types in children in Lamongan Regency based on case data of pneumonia, childhood tuberculosis, non-pneumonia cough, and coughing toddlers. In addition, this study compared the performance of the two clustering algorithms using the Silhouette Score and Davies-Bouldin Index (DBI) evaluation metrics. The results of the study are expected to provide spatial information on the distribution of childhood respiratory diseases and support policy making in the field of regional public health.

II. METHODOLOGY

A. Data Collecting

The data was obtained from the Lamongan District Health Office through the Lamongan Health Profile. The dataset consists of 32 records and 6 parameters related to pediatric respiratory disease cases in Lamongan Regency. Table I below contains the metadata of the dataset used.

TABLE I. DATA ATRIBUTE

ID	Parameters	Description	Type Data
1	Sub-District	Sub-Districts within Lamongan Regency, East Java, Indonesia	String
2	Community Health Center (Puskesmas)	Community health centers under the local government of Lamongan Regency, East Java, Indonesia	String
3	0-14 Childhood Tuberculosis	TB cases reported in children under 15 years old.	Numeric
4	Under-Five Pneumonia	Pneumonia cases in children aged 0–59 months.	Numeric
5	Non-Pneumonia Cough	Cough cases not classified as pneumonia.	Numeric
6	Chough in Toddlers	All cough cases in children under five, regardless of cause.	Numeric

B. Data Preprocessing

Data preprocessing was carried out to prepare the data for the next stage, which is clustering. The first step involved grouping the data based on sub-districts. This was necessary because the available data was categorized by community health centers (puskesmas), and several sub-districts have more than one puskesmas. Since the clustering aimed to identify the distribution of pediatric respiratory disease cases by sub-district, the data was aggregated accordingly. This grouping was performed using the group by function in the Python programming language.

In this stage, data normalization was also performed to adjust the scale of the data values.

Normalization is important because features with varying ranges can negatively affect the performance of the clustering model. The normalization technique used was standardization with Standard Scaler. Standard Scaler transforms the data to have a mean of 0 and a standard deviation of 1.

$$z = \frac{x-\mu}{\sigma} \quad (1)$$

Where :
z : Standard Scaled Value
x : Original Value
μ : Feature Mean
σ : Feature Standard

C. Clustering

Hierarchical clustering is a technique that organizes data into a dendrogram or nested cluster structure. Unlike other approaches, it does not require specifying the number of clusters beforehand. The grouping process produces a dendrogram that illustrates the merging (agglomerative) or splitting (divisive) of clusters [3]. In agglomerative clustering, there are four linkage methods used to calculate distances: single, average, complete, and Ward linkage.

1. Single linkage

   Single linkage, or minimum distance, calculates the distance between two clusters by measuring the shortest distance between any pair of members from the two clusters. This technique is capable of capturing non-elliptical shapes but is sensitive to outliers. The mathematical formulation is presented in Equation (2).

$$D(A,B) = min_{a\in A, b\in B}\, d(a,b) \quad (2)$$

2. Complete linkage

   Complete linkage calculates the maximum distance between any two points from different clusters to measure cluster separation. This method tends to form tighter clusters but may divide naturally shaped groups if they are irregular. The mathematical formulation of complete linkage is presented in Equation (3).

$$D(A,B) = max_{a\in A, b\in B}\, d(a,b) \quad (3)$$

3. Average linkage

   Average linkage, also known as UPGMA, calculates the mean distance between all pairs of points across two clusters. This method strikes a balance between the extremes of single and complete linkage, resulting in more stable clustering outcomes. It is advantageous because it is less sensitive to outliers; however, its main drawback is the relatively more complex computation. The mathematical formulation of average linkage is presented in Equation (4).

$$D(A,B) = \frac{1}{|A||B|}\sum_{a\in A} \quad \sum_{a\in B} \quad d(a,b) \quad (4)$$

4. Ward linkage

   Ward linkage, also known as the minimum variance method, differs slightly from the three previously discussed distance metrics. This technique considers not only the distance between clusters but also the variance within them. Ward linkage merges two clusters in a way that results in the smallest possible increase in total within-cluster variance. It tends to produce homogeneous clusters by minimizing intra-cluster variation. The advantage of Ward linkage lies in its ability to generate compact and similar clusters; however, it may not be suitable when the clusters have irregular shapes or unequal sizes. The mathematical formulation of Ward linkage is presented in Equation (5).

$$\Delta E = \frac{|A||B|}{|A|+|B|}\left|\left|\underline{x}_{\mathrm{A}} - \underline{x}_{\mathrm{B}}\right|\right|^2 \quad (5)$$

Unlike K-Means, which uses the mean as the cluster center, K-Medoids selects an actual data point, enhancing robustness against outliers. Various distance measures, such as Manhattan and Euclidean distances, can be applied in K-Medoids calculations.

1. Euclidean distance
   This metric calculates the direct distance between two points in an n-dimensional space. It is one of the most commonly used distance measures in K-Medoids. The formula represents the Euclidean distance in a plane (6).

$$d(x,y) = \sqrt{\sum_{i=1}^{n} \quad (x_i - y_i)^2} \tag{6}$$

Description :
$d(x.y)$: Euclidean distance between points x and y
n : Number of dimensions (features)
x_i : Value of the i-th feature of point x
y_i : Value of the i-th feature of point y

2. Manhattan distance
   Manhattan distance, also known as the L1-norm or taxicab distance, measures the total distance traveled along the horizontal and vertical axes. This technique is generally more robust to outliers compared to Euclidean distance. The Manhattan distance formula is presented in Equation (7).

$$d(x,y) = \sum_{i=1}^{n} \quad |x_i - y_i| \tag{7}$$

Description :
$d(x,y)$: Manhattan distance between points x and y
n : Number of dimensions
$|x_i - y_i|$: Absolute difference of the i-th feature values

D. Evaluation

Clustering effectiveness can be assessed using the silhouette score, which evaluates how well each point fits within its cluster compared to others. Higher silhouette values, ranging from -1 to 1, signify clearer separation and superior cluster distinction. The silhouette score is computed using the formula presented in Equation (10).

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \tag{10}$$

Description :
s(i) : Silhouette score for data point i
a(i) : Average distance from point i to all other points in the same cluster
b(i) : Average distance from point i to all points in the nearest neighboring cluster

The Davies-Bouldin Index (DBI) gauges clustering quality by measuring intra-cluster compactness and inter-cluster separation. Although DBI lacks a fixed range, lower scores indicate better clustering with tighter, more distinct groups. The formula for DBI calculation is provided in the referenced equation (11).

$$DB = \frac{1}{k}\sum_{i=1}^{k} \quad max_{j \neq i}\left(\frac{s_i + s_j}{d_{ij}}\right) \tag{11}$$

Description :
k : Number of clusters
s_i : Average distance between all points in cluster i and the centroid of cluster i
d_{ij} : Distance between the centroid of cluster i and the centroid of cluster j

The Calinski-Harabasz Index, also known as the Variance Ratio Criterion, is an evaluation method that measures the ratio between inter-cluster dispersion and intra-cluster variation. Higher index values indicate better cluster separation and greater compactness of the formed clusters [8].

The CHI formula is presented in Equation (12).

$$CHI = \frac{Tr(Bk)/(k-1)}{Tr(Wk)/(n-k)} \quad (12)$$

Description :
Tr(Bk) : Total between-cluster variance
Tr(Wk) : Total within-cluster variance
k : Number of clusters
n : Total number of samples

III. RESULTS AND DISCUSSIONS

Lamongan Regency has 27 sub-districts. Clustering analysis of respiratory diseases in children was conducted to determine the distribution of cases based on sub-district areas in Lamongan Regency. In addition, the use of two algorithms in this clustering analysis aims to find out which algorithm can capture the most optimal variation of disease case data. This clustering analysis was executed using computational techniques, namely the python programming language with google collab as the supporting software.

The initial total dataset was 32 data. After clustering based on subdistrict parameters, the dataset shrank to 27 data, corresponding to the number of subdistricts in Lamongan district. The variables used for clustering consisted of variables with numerical data, namely the variables Child Tuberculosis 0-14, Toddler Pneumonia, Cough Not Pneumonia, and Toddler Cough. These variables have different scales, as shown in Table II.

TABLE II. DESCRIPTIVE STATISTICS

	0-14 Childhood Tuberculosis	Under-Five Pneumonia	Non-Pneumonia Cough	Chough in Toddlers
count	27	27	27	27
mean	11,78	104,67	1317,18	1399,81
std	25,58	69,72	1009,52	1023,41
min	0	3	278	282
25%	4	55,50	552,50	687
50%	5	77	1079	1186
75%	11	145,5	1726,5	1818
max	136	263	3970	4047

Based on Table II, it is evident that the four variables under study exhibit a wide range of scales, with values spanning from tens to hundreds and even thousands. Such disparities in scale can negatively affect the performance of clustering algorithms, particularly distance-based methods like Hierarchical Clustering, which are sensitive to the magnitude of numerical values. Therefore, it was necessary to perform data normalization prior to clustering to ensure that each feature contributes equally to the model.

The normalization technique used in this study is Standard Scaler, which is available in the Python programming language via the scikit-learn library, specifically in the *sklearn.preprocessing* module. This technique transforms the data so that each feature has a mean of 0 and a standard deviation of 1. By standardizing the variables, the data becomes more balanced, which enhances the performance and

accuracy of the clustering algorithm.

Following the normalization step, the Hierarchical Clustering algorithm was applied using four different linkage methods: single linkage, complete linkage, average linkage, and ward linkage. Although the dendrograms generated by these methods may appear visually similar at a glance, each method calculates distances in a distinct way, potentially leading to different cluster structures. Figure 1 below presents a dendrogram generated using hierarchical clustering, which illustrates the hierarchical structure of cluster formation among sub-districts based on the similarity of pediatric respiratory disease case patterns.

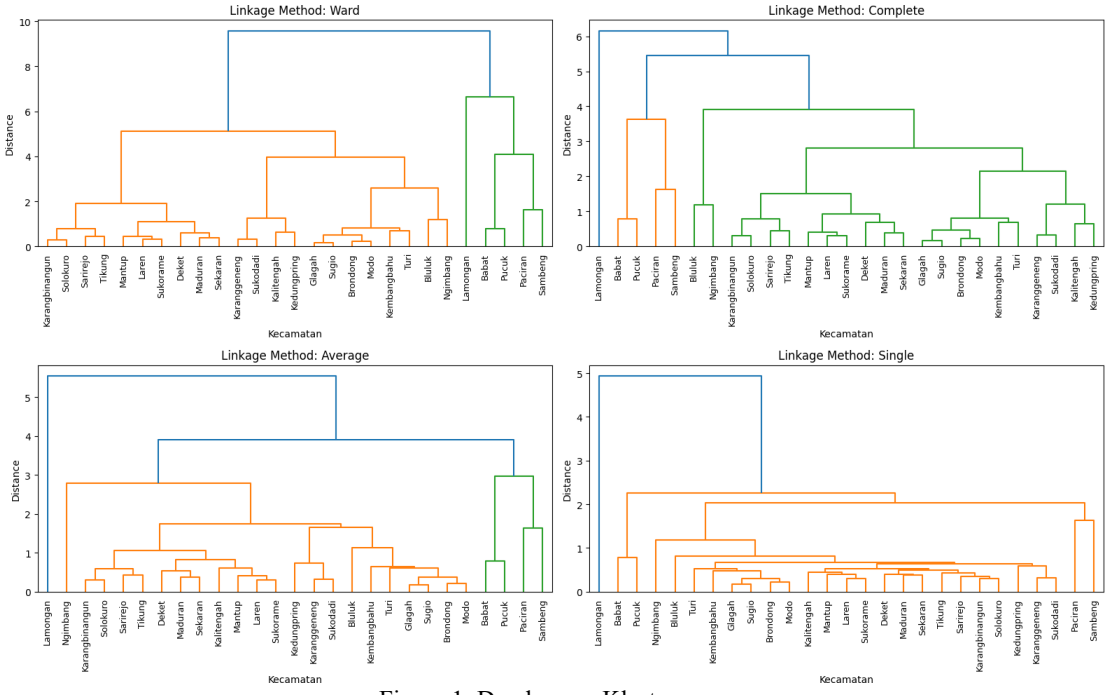

Figure 1. Dendogram Klaster

Based on the results of the dendrogram, it can be seen that the results of hierarchical clustering in the ward linkage distance calculation technique form 3 clusters with a cutting distance of around 4-6. Ward linkage produces clusters with relatively balanced sizes. Meanwhile, in the cluster with complete linkage distance calculation, it can be seen that the large cluster is divided into many small branches. There are also sub-districts that appear as outliers that join last. The average linkage dendrogram appears to have a gradual merging and tends to be flat. There are 3-4 large clusters when cut at a distance of 3. Meanwhile, the single linkage dendrogram forms an elongated chain structure so that no large clusters are formed. The single linkage cluster structure looks very loose. Based on this interpretation, the ward linkage method is a distance calculation metric in hierarchical clustering that produces the most stable clusters compared to other metrics. Ward linkage excels in forming clear clusters and forming balanced clusters. Through dendrogram visualization, it is also known that the optimal number of clusters that can be formed is 3 clusters. Table III below shows the cluster results using the ward linkage method:

TABLE III. WARD LINKAGE METHOD RESULT

Cluster	Sub-districts	Category
1	22 Members: Blubuk, Brondong, Deket, Glagah, Kalitengah, Karangbinangun, Karanggeneng, Kedungpring, Kembangbahu, Laren, Maduran, Mantup, Modo, Ngimbang, Sarirejo, Sekaran, Solokuro, Sugio, Sukodadi, Sukorame, Tikung, Turi	Low
2	4 Members: Babat, Paciran, Pucuk, Sambeng	Medium
3	1 Member: Lamongan	High

In the table of clustering data processing results, it is found that 1 sub-district is included in cluster 3 or high category, then there are 4 sub-districts that are classified in cluster 2 or medium category, and as many as 22 sub-districts are included in cluster 1 or low category. Meanwhile, Figure 2 shows the visualization of clustering in the form of an area distribution map to illustrate the geographical distribution of each cluster more clearly.

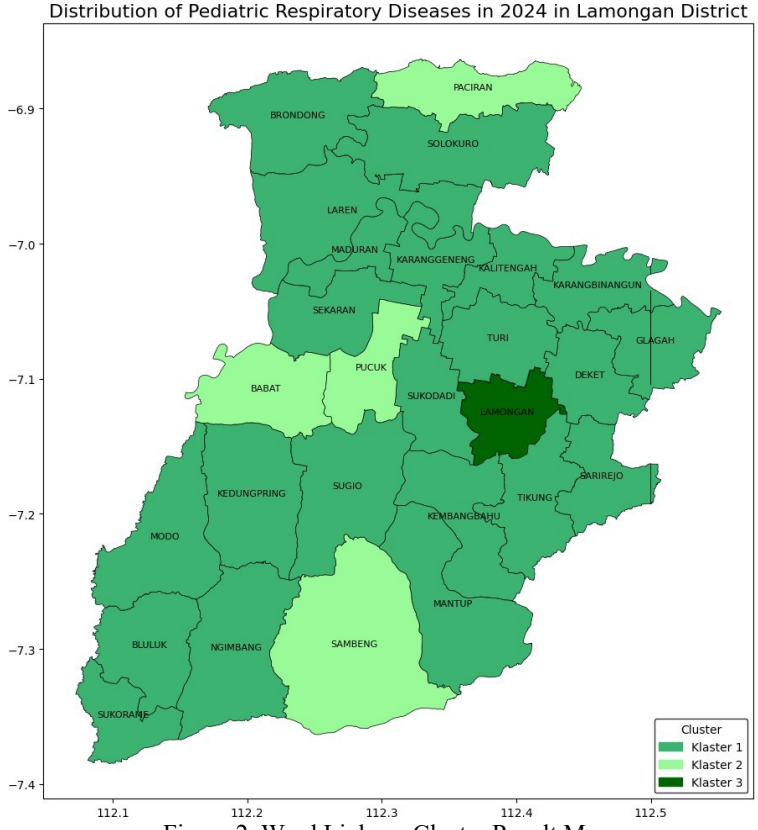

Figure 2. Ward Linkage Cluster Result Map

Furthermore, to determine the performance of the four distance metrics mathematically, an evaluation was conducted. The performance evaluation of the hierarchical clustering method is carried out using four linkage methods, namely Single, Complete, Average, and Ward, which are assessed through three internal metrics namely Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). Table IV below shows the results of the evaluation of distance metrics in hierarchical clustering.

TABLE IV. EVALUASI METODE LINKAGE PADA HIERARCHICAL CLUSTERING

Linkage Method	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
Single	0,542908	0,347034	13,343300
Complete	0,544704	0,588435	20,301826
Average	0,544704	0,588435	20,301826
Ward	0,544704	0,588435	20,301826

Based on the Silhouette Score, the Complete, Average, and Ward methods show identical highest scores of 0.544704, while the Single method is slightly lower with a score of 0.542908. This value indicates that the Complete, Average, and Ward methods have slightly better cluster separation and compactness than the Single method.

On the Davies-Bouldin Index, where the smaller the value means the better the cluster formed, the Single linkage method recorded the lowest score of 0.347034, indicating that the resulting clusters are more compact and separated than other methods. Meanwhile, the other three methods (Complete, Average, and Ward) have identical DBI values of 0.588435, indicating that the cluster separation is not as good as the Single method.

Whereas in the Calinski-Harabasz Index, the highest value of 20.301826 is produced by the Complete, Average, and Ward methods, which indicates that the variance between clusters is much greater than the variance within clusters, which means that cluster separation is more optimal. The Single method produces a lower CHI value of 13.343300, which indicates that this method is less good at maximizing the separation between clusters.

Overall, although Single linkage showed the best results in the Davies-Bouldin Index metric, the other three evaluation metrics (Silhouette and CHI) were superior to the Complete, Average, and Ward methods. In addition, the visualization of the Single Linkage dendrogram resulted in a chain structure, making it less than ideal for use. Since Complete, Average, and Ward produced the same and highest scores on two of the three evaluation metrics, it can be concluded that the Complete, Average, and Ward linkage method is a better choice than the Single linkage method.

Meanwhile, the K-Medoids algorithm compares two distance calculation techniques, namely euclidean and manhattan distance. This algorithm uses a medoid as the cluster center, where the medoid is the centermost object of a cluster, namely the point that has the minimum total distance to its own cluster members. The initial number of clusters set in K-Medoids is 3 clusters, corresponding to the optimal number of clusters formed in the hierarchical clustering algorithm. Performance evaluation of the K-Medoids method was conducted using three evaluation metrics, namely Silhouette Score, Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI). Tests were conducted on two types of distance metrics, namely Euclidean and Manhattan. Table V below shows the evaluation results of both distance calculation metrics in K-Medoids.

TABLE V. EVALUASI K-MEDOIDS BERDASARKAN JARAK

Distance Metrics in K-Medoids	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
Euclidean	0,325567	1,146009	11,374772
Manhattan	0,065092	1,538132	4,808843

Based on the Silhouette Score, which measures how similar the data is to its own cluster compared to other clusters, the Euclidean method shows a higher score of 0.325567, compared to 0.065092 for the Manhattan method. The very low Silhouette score in the Manhattan method indicates that the cluster structure formed using this method is very poor.This value indicates that the cluster separation in the Euclidean method is better and clearer than Manhattan.

Based on the Silhouette Score, which measures how similar the data is to its own cluster compared to other clusters, the Euclidean method shows a higher score of 0.325567, compared to 0.065092 for the Manhattan method. The very low Silhouette score in the Manhattan method indicates that the cluster structure formed using this method is very poor.This value indicates that the cluster separation in the Euclidean method is better and clearer than Manhattan.

Finally, on the Calinski-Harabasz Index, which measures the ratio between inter-cluster variance and within-cluster variance, higher values indicate better cluster quality. The Euclidean method produced a CHI value of 11.374772, much higher than the Manhattan value of 4.808843.

Based on the results of cluster evaluation using 2 different algorithms, namely Hierarchical Clustering and K-Medoids using their respective distance matrix methods. It can be concluded that the Hierarchical Clustering method with Ward linkage is the best clustering method used in this analysis. This method produces the highest Silhouette Score value of 0.5447, which indicates that the clusters formed have good internal cohesiveness and are clearly separated from other clusters. The DBI value of 0.5884 indicates a relatively compact and non-overlapping cluster structure, while the CHI value of 20.3018 indicates an optimal ratio of variance between and within clusters.

In contrast, the K-Medoids method with both Euclidean and Manhattan distances showed much lower performance, with small Silhouette Score and high DBI values, and low CHI. This indicates that the cluster structure formed by K-Medoids is poor in terms of separability and internal consistency. Thus, Hierarchical Clustering with Ward linkage was chosen as the most suitable method in clustering cases of respiratory diseases in children in Lamongan District.

IV. CONCLUSION

Based on the clustering analysis of pediatric respiratory disease cases in Lamongan Regency, the Hierarchical Clustering algorithm using Ward Linkage distance calculation was found to provide the best variation capture among all tested methods, including other distance calculations and the K-Medoids algorithm. This conclusion is supported by evaluation metrics, with a Silhouette Score of 0.544704, a Davies-Bouldin Index of 0.588435, and a Calinski-Harabasz Index of 20.301826. The analysis identified three distinct clusters of pediatric respiratory disease cases. The first cluster includes the sub-districts of Blubuk, Brondong, Deket, Glagah, Kalitengah, Karangbinangun, Karanggeneng, Kedungpring, Kembangbahu, Laren, Maduran, Mantup, Modo, Ngimbang, Sarirejo, Sekaran, Solokuro, Sugio, Sukodadi, Sukorame, Tikung, and Turi. The second cluster consists of the sub-districts of Babat, Paciran, Pucuk, and Sambeng. The third cluster is represented solely by the Lamongan sub-district.

The clustering of pediatric respiratory disease cases provides valuable insights for mapping disease distribution, enabling more targeted and effective public health interventions. Clusters with a high number of cases can be prioritized as the main focus for government health actions, while areas with fewer cases may require lighter interventions. Data-driven strategies allow for more impactful and efficient interventions, particularly in terms of resource allocation and service delivery.

V. REFERENCES

[1] Kementerian Kesehatan Republik Indonesia, *Profil Kesehatan Indonesia Tahun 2022*. Jakarta: Kemenkes RI, 2023. [Online]. Tersedia: https://www.kemkes.go.id/resources/download/pusdatin/profil-kesehatan-indonesia/

[2] D. Restiana, A. Ramadhan, dan S. P. Mahendra, "Implementasi Metode K-Means dan K-Medoids untuk Klasterisasi Penyakit Berdasarkan Gejala Pasien," *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 1, pp. 40–45, 2021, doi: 10.14710/jtsiskom.9.1.2021.40-45.

[3] P.-N. Tan, M. Steinbach, A. Karpatne, dan V. Kumar, *Introduction to Data Mining*, 2nd ed. Boston, MA: Pearson, 2019.

[4] L. Kaufman dan P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley & Sons, 2009.

[5] J. Han, M. Kamber, dan J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA: Morgan Kaufmann, 2012.

[6] T. Hastie, R. Tibshirani, dan J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York: Springer, 2009.
[7] D. L. Davies dan D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.
[8] T. Calinski dan J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics*, vol. 3, no. 1, pp. 1–27, 1974, doi: 10.1080/03610927408827101.
[9] S. Tuhpatussania, S. Erniwati, dan Z. Mutaqin, "Perbandingan metode agglomerative hierarchical clustering dan metode K-Medoids dalam pengelompokan data titik panas kebakaran hutan di Indonesia," *Journal Computer and Technology*, vol. 2, no. 1, pp. 21–38, Jul. 2024, doi: 10.69916/comtechno.v2i1.146.
[10] G. R. Suraya dan A. W. Wijayanto, "Comparison of Hierarchical Clustering, K Means, K Medoids, and Fuzzy C Means methods in grouping provinces in Indonesia according to the special index for handling stunting," Indonesian Journal of Statistics and Its Applications, vol. 6, no. 2, pp. 180–201, Aug. 2022, doi: 10.29244/ijsa.v6i2p180-201.