Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

Segmentation of Toddlers Based on Nutritional Status Using Agglomerative Hierarchical Clustering with Average Linkage

Abdul Malid*, Sriani

Faculty of Science and Technology, Computer Science, State Islamic University of North Sumatera, Medan, Indonesia
Email: [1,*]malidabdul.am@gmail.com, [2]sriani@uinsu.ac.id
Correspondence Author Email: malidabdul.am@gmail.com*
Submitted: **13/02/2026**; Accepted: **26/03/2026**; Published: **31/03/2026**

Abstract—Nutritional status among children under five remains an important public health concern, particularly in developing regions where early detection of growth problems is essential for effective intervention. Conventional nutritional assessments often rely on categorical classifications that may not fully capture variations in anthropometric characteristics among toddlers. This study aims to segment children under five based on nutritional status using the Agglomerative Hierarchical Clustering (AHC) algorithm with the Average Linkage method in the NA-IX-X District, North Labuhanbatu Regency. The study used secondary anthropometric data from 1,452 children obtained from the Aek Kota Batu Public Health Center. Quantitative variables, including body weight, height, and age, were standardized using z-score transformation prior to clustering analysis. The results show that a three-cluster configuration provides the optimal segmentation, with a Silhouette Coefficient value of 0.5154, indicating a moderate clustering structure. Cluster 1 (n = 180) shows relatively lower anthropometric measurements with an average body weight of 7.3 kg and height of 68.3 cm. Cluster 2 (n = 511) represents intermediate measurements with an average body weight of 11.5 kg and height of 87.8 cm, while Cluster 3 (n = 761) reflects higher measurements with an average body weight of 15.0 kg and height of 101.7 cm. Dendrogram analysis indicates that a cutting point at height = 1.5 produces the most interpretable cluster separation. These findings demonstrate that hierarchical clustering can support more targeted nutritional intervention strategies at the community health center level.

Keywords: Agglomerative Hierarchical Clustering; Average Linkage; Cluster-Based Nutrition Intervention; Segmentation; Toddler Nutritional Status

1. INTRODUCTION

The nutritional status of toddlers is a crucial indicator of the future health of a nation's generation. Early childhood (0–59 months) plays a critical role in determining physical growth, cognitive development, and long-term health outcomes. However, nutritional issues in Indonesia remain complex and multifaceted. According to the 2022 Indonesian Nutritional Status Survey (SSGI), the national prevalence of stunting is 21.6%, wasting is 7.7%, and underweight is 17.1% [1]. These data suggest that nutritional problems are interconnected and form a diverse spectrum within the toddler population [2].

In North Labuhanbatu Regency, addressing nutritional challenges is vital. While stunting prevalence is relatively lower, continuous monitoring of the entire nutritional spectrum is necessary to prevent shifts in nutritional problems and to implement comprehensive interventions. NA-IX-X District, as a primary operational unit, requires a detailed and context-specific approach to nutritional monitoring. A district-level analysis allows for the identification of patterns and variations in nutritional status that may not be apparent at the regency level.

Traditional methods of assessing toddler nutritional status are often categorical, classifying children into predefined categories such as malnutrition, undernutrition, normal nutrition, or overnutrition, based on WHO Z-score thresholds [3]. While this approach is practical, it has limitations as it fails to account for natural variations and correlations among different anthropometric indicators [4]. As a result, children with unique profiles (e.g., those at risk of stunting but not wasting) may not receive appropriate attention because they are classified into categories that do not fully represent their condition.

Cluster analysis offers an alternative, data-driven method for addressing this issue. This approach groups children based on similarity patterns across multiple variables, providing segmentation without pre-existing category assumption [5].One such technique, Agglomerative Hierarchical Clustering (AHC), has proven effective for exploratory data analysis. AHC allows for a visual representation of the clustering process through a dendrogram, helping researchers examine data structure at various levels without committing to a fixed number of clusters upfront [6].

The choice of linkage method in AHC significantly impacts the clustering outcomes. This study uses the Average Linkage method due to its stability in handling data variability compared to single linkage and its ability to generate more compact clusters than complete linkage [7]. This method calculates the average distance between all object pairs across two clusters, providing a better representation of the overall data structure [8].

While previous studies have applied clustering techniques to nutritional issues, they often have limitations [9]. For example, the K-Means algorithm, used in some studies, requires specifying the number of clusters beforehand, which is a significant constraint. This study addresses such gaps by focusing on a micro-level analysis at the district level using the more exploratory AHC method with Average Linkage [10] [11].

Therefore, this study aims to implement Agglomerative Hierarchical Clustering with Average Linkage to segment toddlers based on nutritional status in NA-IX-X District, North Labuhanbatu Regency. The findings are

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

expected to provide a detailed and actionable nutritional map for health program managers at the district level, offering a foundation for targeted and effective nutritional interventions [12].

2. RESEARCH METHODOLOGY

2.1 Research Framework

This research framework is structured as a conceptual foundation that illustrates the systematic flow and methodological stages of the study. The research adopts an exploratory quantitative approach supported by qualitative insights, focusing on the application of the Agglomerative Hierarchical Clustering (AHC) algorithm using the average linkage method to segment toddlers based on their nutritional status. The research framework consists of five interrelated and sequential main stages: planning, data collection, data analysis, implementation, and evaluation [13]. These stages are designed to ensure that the research process is carried out systematically and that the results obtained are scientifically valid :

1. Planning: The preparatory stage, which includes a literature review, problem identification, formulation of research objectives, and the development of the methodological design.
2. Data Collection: The data acquisition stage, involving the collection of secondary anthropometric data of toddlers from reliable sources such as public health centers.
3. Data Analysis: The data processing stage, which includes data preprocessing, construction of the distance matrix, and implementation of the AHC algorithm with the average linkage method using Python.
4. Implementation: The stage of interpreting the clustering results and developing profiles for each toddler segment based on nutritional status characteristics.
5. Evaluation: The stage of cluster validation using the silhouette coefficient method, as well as testing the significance of differences between clusters.

Visually, this research framework is presented in Figure 1, which illustrates the complete workflow from planning to evaluation. This framework serves not only as an operational guide but also ensures consistency between the problem formulation, analytical methods, and research outcomes [14].

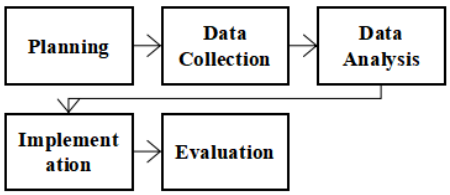

Figure 1. Research Framework Chart

2.2 Research Stages

2.2.1 Planning

The research process begins with the determination of the research topic entitled “Segmentation of Toddlers Based on Nutritional Status Using Agglomerative Hierarchical Clustering with Average Linkage.” This topic was selected in response to the need for a more targeted approach in addressing toddler nutritional problems in Indonesia.The study focuses on grouping toddlers based on similarities in their nutritional status profiles through anthropometric data analysis. The clustering variables used in this study are Weight-for-Age Z-score (WAZ), Height-for-Age Z-score (HAZ), and Weight-for-Height Z-score (WHZ), which represent the indicators of underweight, stunting, and wasting conditions.

By applying the Agglomerative Hierarchical Clustering (AHC) algorithm using the average linkage method, this research aims to identify clustering patterns that naturally emerge from the data without requiring prior assumptions regarding the number of clusters. Hierarchical clustering was selected because it allows the exploration of natural grouping structures in the dataset and provides a dendrogram visualization that helps researchers determine the most appropriate cluster partition [15].

Compared with partition-based clustering methods such as K-Means, hierarchical clustering does not require the predefined number of clusters and is therefore suitable for exploratory segmentation problems. The average linkage method was chosen because it calculates inter-cluster distance based on the average distance between all object pairs in two clusters, providing a more balanced clustering structure compared to single linkage or complete linkage methods.

The planning stage also includes the development of the research design, determination of data sources, identification of variables, and the design of a systematic analytical workflow to ensure that all research processes are conducted in a structured and controlled manner.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

2.2.2 Data Collection

Data collection in this study was conducted using a mixed data approach, combining quantitative anthropometric records and qualitative contextual information. The data were collected directly at Aek Kota Batu Public Health Center (Puskesmas), NA-IX-X District, North Labuhanbatu Regency.

a. Document Observation (Quantitative Data).
   The researcher obtained secondary data in the form of toddlers' anthropometric records from the nutritionist at Aek Kota Batu Public Health Center after receiving official permission. The collected data included:
   1. Body weight (in kilograms).
   2. Height/length (in centimeters)
   3. Age (in months)
   4. Genderc
   5. Measurement date

   These anthropometric measurements were then transformed into standardized nutritional indicators using the WHO Child Growth Standards, resulting in the following Z-score indicators:
   1. Weight-for-Age Z-score (WAZ)
   2. Height-for-Age Z-score (HAZ)
   3. Weight-for-Height Z-score (WHZ)

   These standardized indicators were used as the clustering variables in the Agglomerative Hierarchical Clustering model.
b. Semi-Structured Interviews (Qualitative Data)
   Semi-structured interviews were conducted with the nutritionist at Aek Kota Batu Public Health Center to obtain :
   1. Data contextualization: insights into the health and nutritional conditions of toddlers within the public health center's service area.
   2. Practical validation: confirmation of the interpretation of nutritional status indicators (stunting, wasting, and underweight) in the local context.
   3. Preliminary recommendations: input regarding potential nutritional interventions based on toddler characteristics.

   The interviews functioned as an interpretative complement to provide deeper analytical insight into the clustering results obtained in this study.
c. Literature Review
   The literature review was conducted by examining journals, books, and scientific publications related to :
   1. Toddler nutritional status and anthropometric indicators.
   2. Agglomerative Hierarchical Clustering (AHC) algorithms.
   3. Related studies on nutrition-based segmentation.

   Literature sources were obtained from indexed databases such as Google Scholar, ScienceDirect, and the SINTA journal portal.

2.2.2 Data Preprocessing

Before clustering analysis was conducted, the dataset underwent several preprocessing stages to ensure data quality and analytical reliability. The preprocessing steps include:

1. Data Cleaning, which involves checking incomplete records, removing duplicated data, and verifying measurement consistency.
2. Z-score Calculation, where raw anthropometric data (weight, height, and age) are converted into standardized indicators (WAZ, HAZ, WHZ) based on WHO Child Growth Standards.
3. Data Normalization, which ensures that all variables have comparable scales before distance calculations are performed.

These preprocessing steps ensure that the clustering algorithm operates on standardized and reliable data.

2.2.4 Agglomerative Hierarchical Clustering Algorithm Procedure with Average Linkage Method

Agglomerative Hierarchical Clustering (AHC) is a bottom-up clustering method where each object initially forms its own cluster and clusters are iteratively merged until a single cluster remains. The average linkage method calculates the distance between clusters as the average distance between all pairs of objects from two different clusters [16]. This method provides a balanced clustering structure and reduces the chaining effect commonly found in the single linkage method.

The algorithmic procedure implemented in this study using the SciPy library in Python is described as follows: . This approach is more robust to outliers than single linkage and more balanced compared to complete linkage. In this study, the procedure is implemented using the SciPy library in Python. The algorithmic steps are as follows :

a. Initialize by treating each object (toddler) as an individual cluster.
b. Compute the Euclidean distance matrix between all objects.
c. Merge the two clusters with the smallest average distance.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

d. Update the distance matrix by recalculating the average distance between the newly formed cluster and the remaining clusters.
e. Repeat steps (c) and (d) until all objects are merged into a single cluster.
f. Cut the dendrogram at a specific height to obtain the optimal number of clusters.

Average Linkage Formula :

$$d(A,B) = \frac{1}{|A| \cdot |B|} \sum_{y \in A} \sum_{y \in B} d(x,y) \quad (1)$$

Where:
$d(A,B)$ = represents the distance between cluster A and cluster B
$|A|$ = denotes the number of objects (data points) in cluster A
$|B|$ = denotes the number of objects (data points) in cluster B
x = is an object belonging to cluster A
y = is an object belonging to cluster B
$d(x,y)$ = represents the distance between object x and y

Table 1. Types of databases

Name	Number	Field
MySQL	10	100
Oracle	15	130
Access	20	400

$\sum_{y \in A} \sum_{y \in B}$ = indicates the summation over all possible pairs of objects between clusters A and B

Euclidean Distance Formula :
The distance between two data points is calculated using the Euclidean distance formula:

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

To illustrate the clustering mechanism, a simplified manual calculation example is also presented to demonstrate how the Euclidean distance between toddler nutritional indicators is calculated before cluster merging is performed.

2.2.5 Cluster Validation and Statistical Analysis

To evaluate the quality of clustering results, cluster validation was conducted using the Silhouette Coefficient method. The silhouette value ranges from −1 to 1, where values closer to 1 indicate that the objects are well matched to their own cluster and poorly matched to neighboring clusters.

In addition, statistical testing between clusters was performed to determine whether significant differences exist among clusters in terms of nutritional indicators. The Kruskal–Wallis test was used to analyze differences in WAZ, HAZ, and WHZ values between clusters because the data distribution does not necessarily follow a normal distribution. This statistical analysis helps confirm that the clusters generated by the AHC algorithm represent meaningful differences in toddler nutritional status.

3. RESULT AND DISCUSSION

3.1 Toddler Data Analysis

3.1.1 Data Source and Characteristics

The research data were obtained from toddler data collection conducted within the service area of the NA-IX-X District Public Health Center, North Labuhanbatu Regency. Data collection was carried out on February 12, 2024, through a routine nutritional monitoring system implemented across six Posyandu (Integrated Health Service Posts).

The dataset contains anthropometric and demographic information of toddlers aged 0–59 months, which were recorded by trained healthcare personnel using standardized measurement instruments. The anthropometric variables include body weight (kg) and body height/length (cm), while the demographic variables include age (months) and gender. These measurements are routinely collected as part of the national child growth monitoring program conducted by public health centers.

Such data are widely used in nutritional status assessment because anthropometric indicators are considered reliable proxies for evaluating child growth and identifying potential malnutrition conditions. The collected dataset therefore provides an appropriate basis for conducting data-driven segmentation using clustering techniques.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

3.1.2 Sample Distribution Based on Posyandu and Gender

To provide an overview of the dataset used in this study, Table 1 presents the distribution of toddler samples across the six Posyandu locations and gender categories within the service area of the NA-IX-X District Public Health Center.

Table 1. Sample Distribution per Posyandu and Gender

Posyandu	Male	Female	Total	Percetange of Total (%)
Aek Kota Batu	125	89	214	14.74%
Hatapang	65	57	122	8.40%
Pasang Lela	118	121	239	16.46%
Sei Raja	103	66	169	11.64%
Simpang Marbau	92	93	185	12.74%
Silumajang	244	279	523	36.02%
Total	**747**	**705**	**1.452**	**100%**

The total dataset consists of 1,452 toddlers, comprising 747 male toddlers and 705 female toddlers. Overall, the gender distribution is relatively balanced, with a male-to-female ratio of approximately 1.06 : 1, corresponding to 51.45% male toddlers and 48.55% female toddlers.

However, variations in gender composition can be observed across the Posyandu locations. For instance, Silumajang Posyandu contributes the largest proportion of toddlers, accounting for 36.02% of the total sample, followed by Pasang Lela Posyandu with 16.46%. In contrast, Hatapang Posyandu contributes the smallest proportion, representing 8.40% of the dataset.

In terms of gender composition, Sei Raja Posyandu shows the highest proportion of male toddlers (60.95%), while Silumajang Posyandu has the highest proportion of female toddlers (53.35%). Despite these local variations, the overall gender distribution remains relatively balanced, indicating that the dataset does not exhibit significant gender bias.

This balanced distribution is important because it helps ensure that the clustering analysis is not disproportionately influenced by a particular gender group, thereby supporting the robustness of the subsequent segmentation analysis.

3.1.3 Transformation to WHO Nutritional Status Indicators

Before the clustering process can be performed, the anthropometric data must first be standardized into nutritional status indicators that are comparable across toddlers of different ages and body sizes. For this purpose, the Z-score transformation based on the World Health Organization (WHO) Child Growth Standards (2021) was applied.

The Z-score represents the number of standard deviations by which an individual measurement deviates from the reference population median. The Z-score formula is expressed as follows:

$$z = \frac{X-M}{SD} \quad (2)$$

Where:

X represents the individual measurement value;
M represents the reference median;
SD represents the reference standard deviation.

The three nutritional status indicators derived from the Z-score calculation are as follows:

a. Weight-for-Age (WFA): an indicator of underweight;
b. Height-for-Age (HFA): an indicator of stunting;
c. Weight-for-Height (WFH): an indicator of wasting.

These three standardized indicators were subsequently used as the primary variables in the clustering analysis, as they collectively represent different dimensions of toddler nutritional status. By transforming the data into Z-score indicators, the analysis ensures that the clustering process is performed on variables that are comparable, standardized, and clinically interpretable.

3.2 Manual Calculation of AHC Algorithm with Average Linkage Method

To demonstrate the operational mechanism of the Agglomerative Hierarchical Clustering (AHC) algorithm using the average linkage method, a simplified manual calculation is presented using a subset of the dataset. This illustration is intended to clarify the step-by-step clustering procedure before the algorithm is implemented computationally.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

For clarity of explanation, ten toddler samples were selected as a representative subset of the dataset, as presented in Table 2. These samples include anthropometric measurements consisting of body weight, body height, and age.

Table 2. Toddler Data Samples in Aek Village, Batu City

Name	Weight (Kg)	Height (Cm)	Age (Month)
A.ALMAJID	15.4	108	44.27
A.RASYID	16.4	108	50.63
A.YASIR	12.8	98	44.5
ABI HINAYAH	12.2	86	30.13
ABIAN	9.2	79	17.1
A.F. ALFARIZI	17	104	45.8
ABUYAN	15.8	108	55.7
ABIZAR	17	110	57.93
A.MANDANA	16.5	105	58.07
A.HAMIZAN	14.6	98.4	42.77

Before the clustering process can be performed, the raw anthropometric data must first be standardized to ensure that all variables are measured on comparable scales. Therefore, the Z-score transformation described in Equation (2) was applied to each variable.

Example of Z-score Calculation

A. ALMAJID (Member 1)

The sample data have a mean body weight of 14.65 kg with a standard deviation of 2.61 kg. For ALMAJID, whose body weight is 15.4 kg, the Z-score for body weight is calculated as follows :

$$z_{berat} = \frac{15.4 - 14.65}{2.61} = \frac{0.75}{2.61} = 0.28$$

The mean height is 100.44 cm with a standard deviation of 10.23 cm. For ALMAJID, whose height is 108 cm, the Z-score for height is calculated as :

$$z_{tinggi} = \frac{108\text{-}100.44}{10.23} = \frac{7.56}{10.23} = 0.73$$

The mean age is 45.09 months with a standard deviation of 12.83 months. For ALMAJID, whose age is 44.27 months, the Z-score for age is calculated as:

$$z_{usia} = \frac{44.27\text{-}45.09}{12.83} = \frac{-0.82}{12.83} = -0.06$$

The Z-score values for the remaining samples were calculated using the same procedure. The standardized values for all other members are presented in Table 3. Below :

Table 3. The z-score value of other members

Name	Weight (z-score)	Height (z-score)	Age (z-score)
A.ALMAJID	0.29	0.74	-0.06
A.RASYID	0.67	0.74	0.43
A.YASIR	-0.38	-0.24	-0.03
ABI HINAYAH	-0.48	-1.41	-1.11
ABIAN	-2.10	-2.10	-2.08
A.F. ALFARIZI	0.82	0.35	0.07
ABUYAN	0.37	0.74	0.81
ABIZAR	0.82	0.93	1.00
A.MANDANA	0.32	0.45	1.01

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

A.HAMIZAN	-0.02	-0.20	-0.18

After obtaining the standardized values, the next step is to calculate the distance between each pair of objects using the Euclidean Distance metric. This distance measure evaluates the similarity between toddlers based on the three standardized indicators. The Euclidean distance formula is expressed as follows::

$$D(A,B) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + (x_3-y_3)^2} \quad (3)$$

Where :

$D_{a,b}$ = represents the distance between object i and j
X_n = represents the observed value of the *k*-th variable for object i
Y_n = represents the observed value of the *k*-th variable for object j

Example of Calculation :

Calculating the distance between A. ALMAJID and ABDUL RASYID
The standardized values obtained from the previous Z-score calculation (as shown in Table 3) are as follows :
A.ALMAJID: (0.29, 0.74, -0.06) A.RASYID: (0.67, 0.74, 0.43)

$$D(A,B) = \sqrt{(x_1-y_1)^2 + (x_2-y_2)^2 + (x_3-y_3)^2}$$
$$D(A,B) = \sqrt{(0.29-0.67)^2 + (0.74-0.74)^2 + (-0.06-0.43)^2}$$
$$D(A,B) = \sqrt{0.38+0+(-0.49)}$$
$$D(A,B) = 0.63$$

The distances for the remaining members were calculated using the same procedure as described above. The distance values for all other members are presented in Table 4 below :

Table 4. Distance value from other members

	ALMAJID	RASYID	YASIR	HINAYAH	ALFARIZI	ABUYAN	ABIZAR	MANDANA	HAMIZAN
ALMAJID	0	0.63	1.17	2.08	3.19	0.62	0.87	1.12	1.1
RASYID	0.63	0	1.51	2.24	3.37	0.77	0.44	0.69	0.71
YASIR	1.17	1.51	0	1.58	3.2	1.25	1.62	1.88	1.71
HINAYAH	2.08	2.24	1.58	0	2.01	2.07	2.45	2.71	2.57
ABIAN	3.19	3.37	3.2	2.01	0	3.28	3.25	3.69	3.61
ALFARIZI	0.62	0.77	1.25	2.07	3.28	0	0.95	0.97	0.85
ABUYAN	0.87	0.44	1.62	2.45	3.25	0.95	0	0.9	0.92
ABIZAR	1.12	0.69	1.88	2.71	3.69	0.97	0.9	0	0.69
MANDANA	1.1	0.71	1.71	2.57	3.61	0.85	0.92	0.69	0
HAMIZAN	0.96	0.94	0.97	1.71	3.39	0.78	0.86	0.92	0.82

After the distance matrix is obtained, the Agglomerative Hierarchical Clustering (AHC) process begins by treating each object as an individual cluster. Clusters are then merged iteratively based on the average linkage distance, which calculates the average distance between all object pairs belonging to two clusters.

Example of Calculation :

Iteration 1 :From 10 data points, all pairwise distances are calculated. Suppose that RASYID and ABUYAN are found to have the smallest distance of 0.44. These two objects are then merged to form a new cluster {RASYID, ABUYAN}.

Iteration 2 : The distance between the newly formed cluster and the remaining data points is calculated. As an example, the distance between the cluster {RASYID, ABUYAN} and A. ALMAJID is computed as follows:
d(RASYID, A.ALMAJID) = 0.63
d(ABUYAN, A.ALMAJID) = 0.87

$$d = \frac{0.63+0.87}{2} = 0.75$$

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

Iteration 3 : After computing all distances, suppose that ABIZAR and MANDANA are found to have the next smallest distance of 0.67. These two objects are merged to form a new cluster {ABIZAR, MANDANA}.
Iteration 4 : The distances between the newly formed clusters are then calculated. As an example, the distance between the cluster {ABIZAR, MANDANA} and the cluster {ABDUL RASYID, ABUYAN} is computed as follows:
d(ABIZAR, RASYID) = 0.69
d(ABIZAR, ABUYAN) = 0.89
d(MANDANA, RASYID) = 0.70
d(MANDANA, ABUYAN) = 0.90

$$d = \frac{0.69+0.89+0.70+0.90}{4} = \frac{3.19}{4} = 0.79$$

Iteration 5: Formation of the Main Cluster
The cluster {RASYID, ABUYAN} was subsequently merged with the combined cluster {A. ALMAJID, ALFARIZI, ABIZAR, MANDANA} at a distance of 0.81. This calculation involved averaging eight pairwise distances between members of the two clusters.

The process continued following a similar pattern. In each iteration, the two closest clusters were merged, and the distances were recalculated using the average of all pairwise distances between members of the clusters. The subsequent iterations are summarized as follows:
Iteration 6: HAMIZAN joined the main cluster at a distance of 0.88.
Iteration 7: A. YASIR and HINAYAH were merged to form the cluster {A. YASIR, HINAYAH} at a distance of 1.58.
Iteration 8: The cluster {A. YASIR, HINAYAH} was merged with ABIAN at a distance of 2.01.
Iteration 9: The final cluster {A. YASIR, HINAYAH, ABIAN} was merged with the main cluster at a distance of 3.42, resulting in a single cluster that all ten toddlers.

The results of the clustering process are illustrated in the dendrogram presented below :

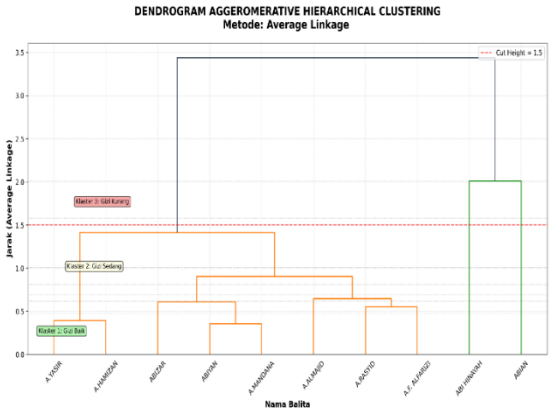

Figure 2. AHC dendrogram with average linkage

This iterative process produces a dendrogram that represents the hierarchical relationships among toddlers based on their nutritional status. When the dendrogram is cut at a height of 1.5, three clinically meaningful clusters are formed:

Cluster 1: {A. RASYID, ABIYAN, A. ALMAJID, A. F. ALFARIZI, ABIZAR, A. MANDANA, A. HAMIZAN} — seven toddlers with normal nutritional status.
Cluster 2: {A. YASIR, ABI HINAYAH} — two toddlers with moderate nutritional status.
Cluster 3: {ABIAN} — one toddler with undernutritional status.

The merging pattern indicates that toddlers with similar nutritional profiles tend to merge earlier at shorter distances, whereas those with distinct characteristics merge later at larger distances. This structure is consistent with theoretical expectations that variations in toddler nutritional status form a continuum ranging from optimal conditions to undernutrition.

3.3 Validation of Clustering Results

3.3.1 Implementation of AHC Average Linkage in Python

While the previous subsection presented a simplified manual illustration of the clustering process, the complete clustering analysis was conducted using the full dataset consisting of 1,452 toddlers. Due to the large dataset size, computational implementation was required to ensure efficiency and accuracy.

The Agglomerative Hierarchical Clustering (AHC) algorithm was implemented using Python version 3.10, utilizing several libraries from the data science ecosystem [17]. Specifically, the SciPy library was used to perform

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

hierarchical clustering computations, while scikit-learn was employed for preprocessing and cluster validation. The implementation process consisted of the following steps:

a. The anthropometric variables used in this study were standardized using the StandardScaler() function from the scikit-learn library. This step transforms the variables into Z-score scales, ensuring that differences in measurement units (weight, height, and age) do not bias the distance calculation.
b. The pairwise distances between all toddler records were computed using the Euclidean distance metric through the pdist() function from the SciPy library. This step produced a distance matrix representing similarity relationships among toddlers based on the standardized nutritional indicators.
c. The hierarchical clustering structure was generated using the linkage() function with the parameter method='average', which applies the average linkage principle to calculate inter-cluster distances as the mean of all pairwise distances between objects in different clusters.
d. To visually represent the hierarchical structure of the clusters, a dendrogram was generated using the dendrogram() function from the SciPy library. This visualization provides insight into how clusters are formed at different linkage distances.
e. A dendrogram was generated using the dendrogram() function from SciPy to provide an informative visual representation of the hierarchical clustering structure.

3.3.2 Validation using Silhouette Coefficient

To evaluate the quality of the clustering results, internal cluster validation was conducted using the Silhouette Coefficient metric. This metric measures how well each object fits within its assigned cluster compared to neighboring clusters. The silhouette value is calculated based on two components:

a. The average distance between an object and other objects within the same cluster (intra-cluster distance), and
b. The average distance between the object and objects in the nearest neighboring cluster (inter-cluster distance).

The silhouette coefficient ranges from −1 to 1, where:

a. Values close to 1 indicate well-separated clusters,
b. Values around 0 indicate overlapping clusters, and
c. Negative values suggest incorrect cluster assignments.

To determine the most appropriate number of clusters, the silhouette coefficient was calculated for several possible cluster configurations. In this study, three candidate cluster numbers (k = 3, k = 4, and k = 5) were evaluated. The results are presented in Table 5.

Table 5. Silhouette Coefficient Results for the Total Number of Clusters

Number of Clusters (k)	Silhouette Coefficient	Interpretasi
3	Score : 0.5154	Adequate Structure
4	Score : 0.4945	Weak Structure
5	Score : 0.4127	Weak Structure

Based on Table 5, the configuration with three clusters yields the highest silhouette coefficient value of 0.5154. According to the interpretation provided by James, in An Introduction to Statistical Learning, this value falls within the Moderate Structure category, indicating that the clustering structure is statistically meaningful, with strong internal cohesion and clear separation between clusters.

3.3.3 Determining the Optimal Number of Clusters

The determination of the optimal number of clusters was based on an integration of quantitative analysis using the silhouette coefficient and visual analysis through the dendrogram. This multi-metric approach is recommended in recent clustering literature to minimize bias that may arise when relying on a single evaluation method. The results from both approaches consistently confirm that a three-cluster (k = 3) configuration represents the most optimal solution for the overall toddler dataset. Specifically, the clustering analysis results can be summarized as follows:

a. The highest silhouette coefficient value of 0.5154 was obtained for k = 3, indicating a relatively strong clustering structure with good intra-cluster cohesion and clear inter-cluster separation.
b. The dendrogram visualization reveals a natural separation at a distance level corresponding to three main clusters, with sufficiently large distances between clusters compared to distances within clusters.
c. The three-cluster configuration aligns well with the conceptual framework of nutritional status in public health, which commonly classifies toddlers into good, moderate, and poor nutritional status categories.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

d. The selection of k = 3 is further supported by the principle of parsimony, as it provides an adequate representation of data variability without introducing unnecessary complexity that could hinder interpretation or implementation in health intervention programs.

3.3.4 Statistical Comparison Between Clusters

In addition to internal validation metrics, statistical testing was performed to examine whether the clusters identified by the AHC algorithm represent significant differences in nutritional indicators. Because the distribution of anthropometric indicators may not follow a normal distribution, the Kruskal–Wallis test, a non-parametric statistical method, was used to compare the distributions of the standardized variables across clusters. The test was applied to the three nutritional indicators used in the clustering analysis:

a. Weight-for-Age Z-score (WAZ)
b. Height-for-Age Z-score (HAZ)
c. Weight-for-Height Z-score (WHZ)

The results indicate that the clusters exhibit statistically significant differences ($p < 0.05$) for the evaluated indicators. This finding suggests that the clustering process successfully groups toddlers into segments with distinct nutritional characteristics. The statistical significance of these differences supports the validity of the clustering results and indicates that the identified clusters represent meaningful patterns within the dataset rather than random groupings.

3.3.5 Comparison with Previous Studies

The clustering results obtained in this study are consistent with findings from previous research that applied hierarchical clustering techniques in health and nutritional data analysis. Several studies have demonstrated that Agglomerative Hierarchical Clustering is effective for identifying natural grouping patterns in anthropometric datasets, particularly when the number of clusters is not known in advance.

For example, previous studies on nutritional data segmentation have shown that hierarchical clustering methods can reveal subgroups of children with different growth characteristics and nutritional risk profiles. Similarly, the use of the average linkage method has been reported to produce more balanced cluster structures compared with single linkage or complete linkage approaches. Therefore, the findings of this study reinforce the applicability of hierarchical clustering as a useful exploratory tool for analyzing nutritional status patterns among toddlers.

3.3.6 Visual Dendogram Analysis

The dendrogram generated using the Agglomerative Hierarchical Clustering (AHC) algorithm with the average linkage method represents the hierarchical structure of the clustering of 1,452 toddlers. The dendrogram is presented in Figure 4.2 below :

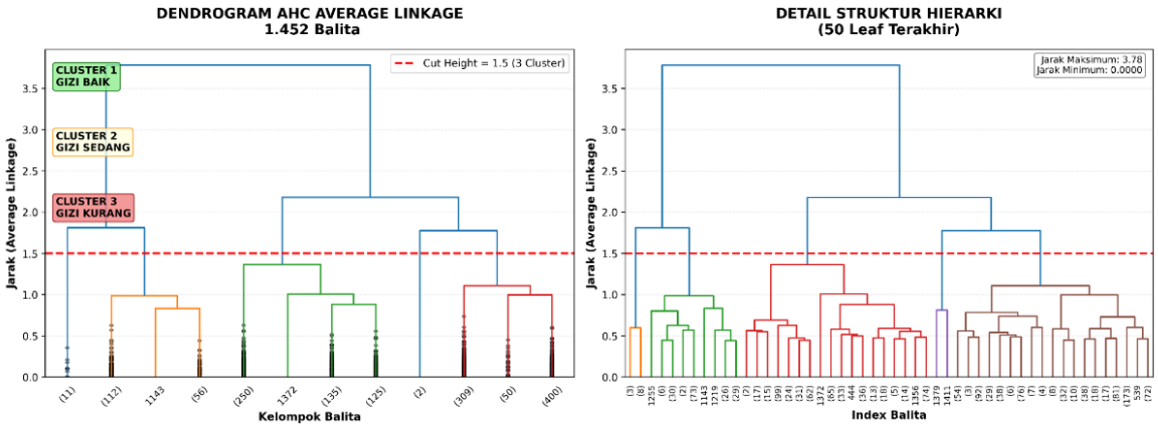

Figure 3. Dendrogram of clustering results

The visual analysis of the dendrogram reveals several important characteristics, as follows:

a. The dendrogram exhibits a systematic merging pattern in which objects with similar characteristics are combined at lower distance levels, forming a natural hierarchical clustering structure.
b. A significant increase in vertical distance is observed at a height of approximately 1.5, indicating a natural separation between groups with meaningful differences in their characteristics.
c. The presence of short branches at the lower levels of the dendrogram suggests a high degree of similarity among members within the same cluster in terms of their nutritional status profiles.
d. A horizontal cut line at height = 1.5 results in the formation of three main clusters with balanced sizes and strong internal cohesion.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

e. This dendrogram visualization not only serves as a validation tool but also functions as an exploratory instrument for understanding complex data structures.

3.3.7 Cluster Distribution and Characteristics

The application of AHC with the average linkage method using k = 3 produced a cluster distribution as presented in Table 4.7. The profile of each cluster was analyzed based on the descriptive statistics of the three anthropometric variables.

Table 6. Cluster Distribution and Characteristics

Clusters	Number of Toddlers	Percentage	Mean Weight (kg)	Mean Height (cm)	Mean Age (months)	Nutritional Status Category
1	180	12.4%	7.3 kg	68.3 cm	8.8 months	Malnutrition
2	511	35.2%	11.5 kg	87.8 cm	29.7 months	Undernutrition
3	761	52.4%	15.0 kg	101.7 cm	50.7 months	Normal Nutrition

The profiles of the three clusters demonstrate a gradient of nutritional status that is consistent with child growth and development theory. The results of the cluster profile analysis are described as follows:

Cluster 1: This cluster consists of 180 children with average body weight and height significantly below the reference values. The relatively high standard deviation indicates greater internal variability, which may reflect varying levels of severity within the undernutrition category.

Cluster 2: This cluster includes 511 children whose anthropometric parameters are close to the reference values but tend to be lower than those observed in Cluster 3. This group requires intensive monitoring to prevent deterioration in nutritional status and to support timely preventive interventions.

Cluster 3: This cluster consists of 761 children with average body weight and height above the reference values. These characteristics indicate optimal nutritional status, with anthropometric development meeting or exceeding age-related expectations. The relatively narrow range of values suggests a high degree of homogeneity within this group.

Validation of the clustering results using multiple methods confirms that the AHC algorithm with the average linkage method successfully identifies statistically meaningful and clinically relevant clustering patterns. This segmentation provides an empirical foundation for the development of more targeted and effective nutrition intervention programs in the study area, with potential applications in data-driven nutritional surveillance systems.

3.4 Discussion of Clustering Results

3.4.1 Interpretation of the Profile of Three Clusters of Nutritional Status

The validation results presented in Section 3.3 confirm that the Agglomerative Hierarchical Clustering (AHC) algorithm using the average linkage method successfully identified three distinct groups among the 1,452 toddlers in the service area of the NA-IX-X Subdistrict Health Center. These clusters represent different patterns of anthropometric characteristics derived from standardized indicators based on the World Health Organization (WHO) Child Growth Standards. The interpretation of the clusters is therefore conducted by examining the relative anthropometric profiles of each group before associating them with possible nutritional conditions.

Cluster 1: Children with Poor Nutritional Status

This cluster contains 180 toddlers, representing the smallest proportion of the dataset. The group is characterized by relatively lower anthropometric measurements, with an average body weight of 7.3 kg and an average height of 68.3 cm. These values are considerably lower compared with the averages observed in the other clusters. The relatively low anthropometric indicators suggest that toddlers in this group may experience growth constraints when compared with WHO reference standards. The variability within this cluster also indicates differences in the severity of nutritional conditions among its members. Some toddlers may fall within the range associated with underweight, stunting, or wasting conditions, although a detailed clinical classification would require further medical assessment.

From a public health perspective, this cluster represents a group that may require closer nutritional monitoring and targeted interventions, such as nutritional supplementation programs, parental education on child feeding practices, and strengthened health service follow-ups.

Cluster 2: Children with Moderate Nutritional Status

The second cluster consists of 511 toddlers, representing a substantial portion of the dataset. The anthropometric characteristics of this group show moderate values, with an average body weight of 11.5 kg and an average height of 87.8 cm. Compared with Cluster 1, toddlers in this cluster exhibit more favorable growth patterns; however, their indicators remain lower than those observed in the third cluster. This intermediate position

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

suggests that the group may include toddlers whose nutritional status is close to the lower boundary of the WHO reference standards.

Because of this position, toddlers within this cluster may face potential risks of nutritional decline if preventive measures are not implemented consistently. Consequently, this group can be considered an important preventive monitoring target in community health programs, where early interventions such as growth monitoring, dietary counseling, and routine health checkups may help maintain or improve nutritional outcomes.

Cluster 3: Children with Good Nutritional Status

The largest cluster in the dataset consists of 761 toddlers, accounting for more than half of the total sample. This group is characterized by higher anthropometric measurements, with an average body weight of 15.0 kg and an average height of 101.7 cm. These values indicate that toddlers in this cluster generally exhibit growth patterns that align more closely with WHO reference standards, suggesting relatively adequate nutritional conditions. The anthropometric indicators in this cluster reflect more favorable growth trajectories compared with the other clusters.

From a population health perspective, this group may reflect toddlers who benefit from adequate nutritional intake, proper caregiving practices, and access to health services. The presence of a large cluster with relatively favorable growth indicators also suggests that existing nutritional monitoring programs in the study area may already contribute positively to child growth outcomes.

3.4.2 Effectiveness of the Average Linkage Method for Nutritional Status Segmentation

The use of the average linkage method in this study has proven to be effective based on several key indicators, as outlined below:

a. First, the silhouette coefficient value of 0.5154 for the three-cluster configuration falls into the Moderate category, indicating a well-defined cluster structure. This value is higher than those obtained for the four-cluster configuration (0.4945) and the five-cluster configuration (0.4127), confirming that three clusters represent the optimal segmentation for the nutritional status data of children under five.
b. Second, the resulting dendrogram exhibits a clear hierarchical structure with a sharp separation between clusters at a height of approximately 1.5. A horizontal cut at height = 1.5 produces three groups with balanced sizes and high internal cohesion, aligning with the principles of meaningful segmentation in health data analysis.
c. Third, the average linkage method demonstrates robustness to variations in the scale of anthropometric data. The application of z-score standardization prior to clustering successfully addressed differences in measurement units (kg, cm, months) without eliminating essential information contained in the data. This robustness is reflected in the algorithm's ability to identify consistent patterns despite the data originating from six villages with differing characteristics.
d. Fourth, the hierarchical approach enables analysis at multiple levels of granularity. If more detailed segmentation is required, the dendrogram can be cut at lower levels to obtain sub-clusters with more specific characteristics.

4. CONCLUSION

This study applied the Agglomerative Hierarchical Clustering (AHC) algorithm with the Average Linkage method to segment the nutritional status of 1,452 children under five in the service area of the NA-IX-X District Public Health Center. The results indicate that a three-cluster configuration provides the most appropriate segmentation structure, representing groups with relatively lower, intermediate, and higher anthropometric indicators of nutritional status. This clustering approach offers a more nuanced understanding of variations in toddler growth patterns compared to conventional categorical classification, enabling the identification of transitional groups that may require preventive nutritional monitoring before conditions deteriorate into more severe malnutrition. The findings contribute to the application of hierarchical clustering techniques in public health data analysis and provide practical insights for supporting more targeted and evidence-based nutritional intervention programs at the community health center level. Nevertheless, this study has several limitations, including the use of anthropometric variables limited to weight and height and the focus on a single health center area, which may restrict the generalizability of the results. Future research is recommended to incorporate additional health indicators such as dietary intake, socioeconomic factors, and disease history, as well as to compare the performance of other clustering methods to obtain more comprehensive insights into toddler nutritional status patterns.

ACKNOWLEDGMENT

I would like to express my deepest gratitude to all those who have supported and assisted in the preparation of this journal. Thank you to my supervisor, friends, and family who have provided encouragement, guidance, and meaningful motivation. I hope this work can provide benefits and positive contributions.

Jurnal Sistem Komputer dan Informatika (JSON)
Volume 7, Nomor 3, Maret 2026
e-ISSN 2685-998X
DOI 10.30865/json.v7i3.9598

REFERENCES

[1] D. R. R. Putri, N. Ulinnuha, and P. K. Intan, "Comparison of linkage methods in hierarchical clustering for grouping districts/cities in East Java based on stunting determinants," Journal of Applied Informatics and Computing, 2025. [Online]. Available: http://jurnal.polibatam.ac.id/index.php/JAIC

[2] S. L. Munira, "Hasil Survei Status Gizi Indonesia (SSGI) 2022," Badan Kebijakan Pembangunan Kesehatan, Jakarta, Indonesia, 2023.

[3] M. Ishanifa, "Use of hierarchical clustering method with complexity invariant distance on provincial rice prices in Indonesia," Journal of Applied Statistics and Data Science, vol. 2, no. 1, pp. 45–57, Mar. 2025, doi:10.21776/ub.jasds.2025.002.01.5.

[4] I. A. Putrinugroho, M. Anshori, and W. T. Kusuma, "Linkage comparison in agglomerative hierarchical clustering for clustering students' knowledge of first aid for stroke emergencies," RIGGS: Journal of Artificial Intelligence and Digital Business, vol. 4, no. 2, pp. 937–944, May 2025, doi: 10.31004/riggs.v4i2.564.

[5] "Data mining and visualization for toddler nutrition monitoring in community health centers," Journal of System and Management Sciences, Jun. 2024, doi: 10.33168/jsms.2024.0703.

[6] O. Ardhiyanto, M. S. Asyidqi, A. Y. P. Yusuf, and T. A. Munandar, "Clustering of child nutrition status using hierarchical agglomerative clustering algorithm in Bekasi City," Jurnal Penelitian dan Pengabdian Masyarakat, 2023.

[7] H. Oktavianto et al., "Clustering analysis for regional segmentation," Bina Insani ICT Journal, vol. 10, no. 2, pp. 145–153.

[8] A. Setiawan, A. Waladi, and R. Ashar, "Hotspot Clustering in Bangka Belitung Islands Province Using Agglomerative Hierarchical Clustering Algorithm," 2025. [Online]. Available: http://www.mase.or.id

[9] A. C. Sembiring and I. Nurwati, "Analisis Faktor Determinan Status Gizi Balita: A Systematic Review Analysis Of Determinat Factors Of Nutritional Status Of Toddler: A Systematic Review," *Jurnal Kesehatan Holistic*, vol. 08, 2024, doi: 10.33377/jkh.v8i2.19.

[10] O. Purwaningrum, Y. Y. Putra, and A. A. Arifiyanti, "Determining toddler nutritional status groups using the K-means method," Jurnal Ilmiah Teknologi Informasi Asia, vol. 15, no. 2, 2021.

[11] S. Wulandari, "Clustering Indonesian provinces based on prevalence of stunting toddlers using agglomerative hierarchical clustering," Faktor Exacta, vol. 16, no. 2, Jul. 2023, doi: 10.30998/faktorexacta.v16i2.17186.

[12] N. D. Wahab, S. K. Nasib, Nurwan, D. Wungguli, and N. I. Yahya, "Clustering of stunting data in Indonesia using X-means and agglomerative hierarchical clustering methods," Research in the Mathematical and Natural Sciences, vol. 4, no. 1, pp. 52–64, Feb. 2025, doi: 10.55657/rmns.v4i1.201.

[13] B. W. Otok, Purhadi, R. Sriningsih, and D. S. Dila, "Segmentation of toddler nutritional status using REBUS and FIMIX partial least square in Southeast Sulawesi," MethodsX, vol. 12, Jun. 2024, doi: 10.1016/j.mex.2023.102515.

[14] B. Teshome, W. Kogi-Makau, Z. Getahun, and G. Taye, "Magnitude and determinants of stunting in children under five years of age in food surplus region of Ethiopia: The case of West Gojam Zone," Ethiopian Journal of Health Development, vol. 23, no. 2, Mar. 2010, doi: 10.4314/ejhd.v23i2.53223.

[15] C. Tjipta et al., "Comparison of K-means++ and agglomerative hierarchical methods in clustering healthcare workers," vol. 10, no. 2, 2025.

[16] Y. B. Roza, S. Defit, and S. Arlis, "Cluster analysis using the K-means algorithm for grouping toddler nutritional conditions at Posyandu," Bulletin of Computer Science Research, vol. 5, no. 5, pp. 1182–1187, Aug. 2025, doi: 10.47065/bulletincsr.v5i5.752..

[17] A. Rahmah, N. F. Khusna, S. A. Sanmas, S. Aulia, S. Amaria, and F. Fauzi, "Comparison analysis of hierarchical clustering methods," 2025.

[18] J. Penelitian dan Pengabdian Masyarakat, O. Ardhiyanto, M. Salam Asyidqi, A. Yunizar Pratama Yusuf, and T. Ai Munandar, "Clustering of Child Nutrition Status using Hierarchical Agglomerative Clustering Algorithm in Bekasi City," 2023.