Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control
Journal homepage: http://kinetik.umm.ac.id
ISSN: 2503-2267
Vol. 11, No. 2, May, Pp. 295-310

Revealing stunting risk patterns through comparative analysis of hierarchical and deep embedded clustering methods

Fifin Ayu Mufarroha*[1], Abdullah Basuki Rahmat[1], Husni[1], Aeri Rachmad[1], Vivin Ayu Lestari[2], Tasya Dwiyanti[1], Malik Maulana[1]
Informatics Engineering Department, University of Trunodjoyo Madura, Bangkalan, Indonesia[1]
Information Technology Dept, State Polytechnic of Malang, Indonesia[2]

Article Info
Keywords:
Stunting, Hierarchical Clustering, Deep Embedded Clustering, Innovation, Big Data.

Article history:
Received: October 29, 2025
Accepted: February 26, 2026
Published: May 01, 2026

Cite:
F. A. Mufarroha, "Revealing Stunting Risk Patterns through Comparative Analysis of Hierarchical and Deep Embedded Clustering Methods", *KINETIK*, vol. 11, no. 2, May 2026.
https://doi.org/10.22219/kinetik.v11i2.2555

*Corresponding author.
Fifin Ayu Mufarroha
E-mail address:
fifin.mufarroha@trunojoyo.ac.id

Abstract
Stunting remains a significant public health issue in Indonesia due to its long-term impact on human resource quality and economic productivity. Despite various intervention programs, disparities in stunting prevalence across regions remain high, particularly in areas characterized by diverse socioeconomic conditions. This study aims to identify regional patterns and group areas based on stunting risk levels using two machine learning approaches: Hierarchical Clustering (HC) and Deep Embedded Clustering (DEC). The data used in this study consist of aggregated toddler measurement data, including the number of toddlers measured, the number of stunting cases, and the percentage of stunting during the 2020–2024 period. The analysis was conducted by comparing the clustering results generated by both methods. The HC method was implemented using the Agglomerative Clustering approach with the Ward linkage criterion, while DEC employed a layered autoencoder architecture optimized using Kullback–Leibler divergence. Cluster quality was evaluated using the Silhouette Score metric. The results show that HC achieved the highest Silhouette Score of 0.5430, while DEC achieved 0.4874, with both methods exhibiting year-to-year performance variation. These findings indicate that HC provides better clustering stability, whereas DEC demonstrates greater adaptability to data complexity and nonlinear patterns. The integration of both methods offers a comprehensive big data–driven health analytics framework, representing an innovative approach for evidence-based decision-making in identifying and addressing stunting-prone regions.

1. Introduction

Stunting remains a critical public health challenge in Indonesia due to its significant impact on human resource quality and economic productivity. Despite various intervention programs, regional disparities in stunting prevalence persist because of diverse socioeconomic conditions [1]. The condition also has long-term consequences for cognitive development and productivity during adulthood. Therefore, addressing stunting constitutes an essential component of sustainable health sector development.

In Indonesia, the prevalence of stunting remains relatively high despite the implementation of various government intervention programs. Based on the results of the 2023 Indonesian Nutritional Status Survey (SSGI), the national stunting prevalence rate was recorded at 21.5%, with considerable variation across regions [2]. The government has implemented various intervention programs, including improving maternal and child nutrition, enhancing sanitation, and strengthening primary healthcare services. However, the reduction in stunting prevalence has not been evenly distributed across all regions [3].

Geographic variation and complex socioeconomic factors indicate that stunting is not solely a health issue but is also influenced by environmental conditions, educational background, and community welfare levels [4]. Therefore, a data-driven analytical approach is needed to uncover hidden patterns and identify stunting-prone regions more accurately. This approach is crucial as a basis for decision-making and the formulation of more targeted intervention policies.

Efforts to map stunting-prone regions in Indonesia have generally focused on presenting descriptive prevalence statistics across administrative areas. While this approach provides a general overview of regional stunting conditions, it does not fully capture patterns of similar risk levels among regions [3]. Most public health reports present data in the form of stunting percentages by district or sub-district without conducting further analytical processes to identify clusters of regions with similar characteristics.

However, identifying these patterns is essential for recognizing groups of regions with similar risk characteristics, thereby enabling interventions to be tailored according to the contextual needs of each area. Although the available

data are relatively simple, such as the number of toddlers measured and the percentage of stunting cases, these data possess substantial potential for further analysis using computational approaches.

One region that presents unique challenges in this context is Madura Island. Based on the 2023 SSGI, districts located on Madura Island continue to exhibit relatively high stunting prevalence compared to the average prevalence in East Java Province [2]. Although the available data are generally limited to the number of toddlers measured and the percentage of stunting cases, this information remains valuable for identifying distribution patterns in high-risk areas. By applying clustering analysis, regions within Madura can be grouped according to similar stunting characteristics, resulting in a more informative risk cluster map.

Advances in computing technology and the increasing availability of health sector data have encouraged the development of new approaches for understanding social phenomena such as stunting. Within the framework of big data analytics, every piece of information, regardless of scale [5], [6], can contribute to building a broader understanding of public health patterns and risks. Big data not only involves large data volumes but also requires the capability to extract meaningful knowledge from complex, heterogeneous, and distributed datasets [7], [8].

In this context, machine learning serves as an important analytical foundation because it enables automatic data processing and the discovery of hidden structures that are difficult to identify using conventional statistical approaches [9]. One relevant technique for uncovering these hidden structures is clustering, an unsupervised learning approach that aims to group objects with similar characteristics into the same cluster. Through clustering analysis, regions can be grouped according to similar levels of stunting risk, thereby generating a more informative spatial risk map for policymaking purposes.

In this study, two clustering methods are comparatively evaluated: Hierarchical Clustering (HC) and Deep Embedded Clustering (DEC). HC is a classical clustering method that forms a hierarchical cluster structures based on similarity distances among regions [10], whereas DEC is a deep learning-based clustering approach that integrates representation learning with simultaneous cluster optimization [11]. These two methods represent two distinct paradigms in big data analytics: a traditional mathematical distance-based approach and a modern nonlinear representation learning approach.

Although clustering techniques have been widely applied in public health studies and regional health mapping, most previous studies primarily relied on conventional algorithms such as K-Means or Hierarchical Clustering without systematically comparing them with deep learning-based clustering approaches [12], [13]. In the context of stunting analysis in Indonesia, existing studies generally focus on descriptive statistics, regression analysis, or single-method clustering approaches, emphasizing prevalence estimation rather than unsupervised pattern discovery [3].

While traditional clustering techniques are effective for identifying similarity structures based on explicit distance metrics, they often assume linear separability within the original feature space and may fail to capture complex nonlinear relationships inherent in heterogeneous regional health data [14]. Recent developments in representation learning indicate that deep clustering frameworks, such as Deep Embedded Clustering (DEC), are capable of uncovering latent structures in multidimensional datasets by simultaneously optimizing feature representation and cluster assignment [11], [14].

However, empirical comparisons between classical distance-based clustering methods and deep learning-based clustering approaches for regional stunting risk analysis remain limited. Consequently, there is still insufficient evidence regarding whether representation learning approaches provide substantial advantages over traditional clustering methods in identifying spatial stunting risk structures.

By comparing the results of HC and DEC, this study aims to demonstrate how advances in machine learning can enrich understanding of stunting risk distribution in Indonesia, particularly in the Madura Island region. The application of these two methods is intended not only to identify stunting-prone areas on Madura Island but also to evaluate the effectiveness of traditional and deep learning-based approaches for spatial public health analysis.

Furthermore, this study emphasizes the importance of integrating data-driven analysis into public policy formulation, where intervention strategies can be designed based on empirical findings generated by computational analysis. The clustering results are expected to provide an initial foundation for understanding regional spatial patterns and supporting the development of more targeted local intervention policies.

2. Research Method

This study employed two unsupervised learning methods: Hierarchical Clustering (HC) and Deep Embedded Clustering (DEC). The primary objective of this research was to identify regional patterns and group areas based on stunting risk levels, as well as to compare the performance of both methods in forming clusters of regions with similar characteristics. The overall research workflow is illustrated in Figure 1.

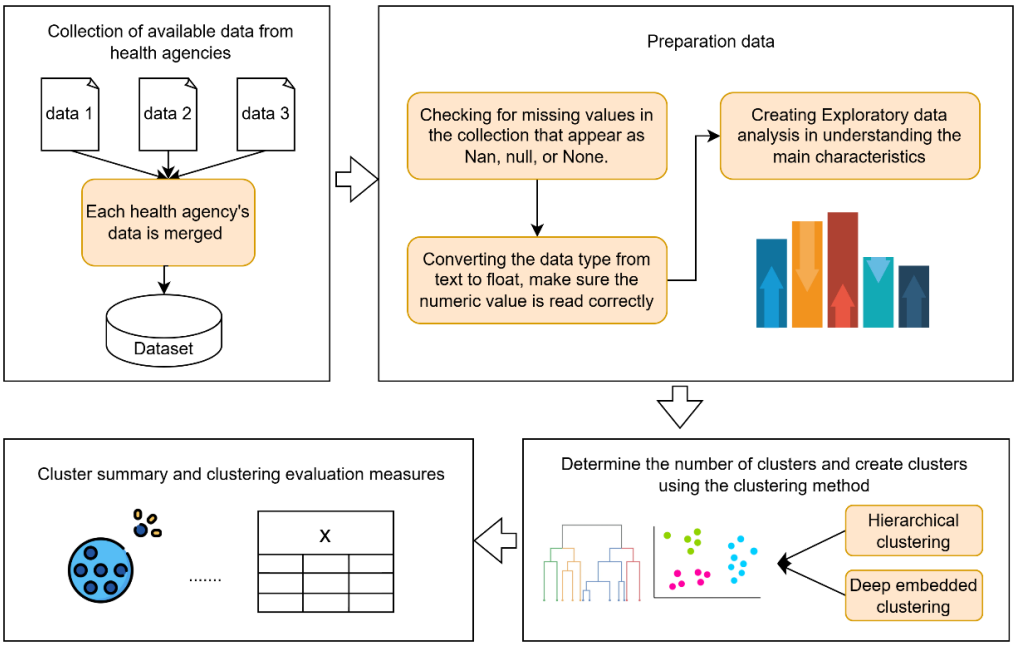


Figure 1. Research Workflow

The data used in this study consisted of secondary data collected from several health agencies located in the Madura Island region, including the Health Offices of Bangkalan Regency, Sampang Regency, and Pamekasan Regency. The data collection process was conducted by integrating datasets from multiple sources into a unified database.

The clustering analysis utilized three primary variables: the number of toddlers measured, the number of stunting cases, and the percentage of stunting prevalence. The first two variables represent the total burden and screening coverage, whereas the stunting percentage reflects the relative risk level. By combining both absolute and relative measurements, the clustering algorithm can distinguish between regions with large population sizes and regions with high stunting risk levels. These variables were selected based on the consistency of data availability across districts and their direct relevance to regional stunting risk characterization.

Subsequently, the data preparation stage was conducted to ensure that the dataset was suitable for computational analysis. Several preprocessing procedures were performed, including checking for missing values, where all data identified as NaN, null, or None were corrected to maintain data consistency. Numeric variables originally stored in text format were then converted into float data types. Furthermore, all variables were standardized to equalize the measurement scale among variables.

Exploratory Data Analysis (EDA) was subsequently conducted to understand the distribution of the dataset. During the clustering stage, the optimal number of clusters was determined, followed by the identification of similarity relationships among regions using the HC and DEC methods. Both methods were then compared in terms of cluster structure, the number of groups formed, and the consistency of clustering results based on the Silhouette Score evaluation metric.

The determination of parameters such as latent dimension, number of hidden units, and number of training epochs for the DEC method is presented in Table 1.

Table 1. Dataset Structure

Parameter	Value	Explanation
Input dimension	3	Three numerical input features
Latent dimension	2	Reduced-dimensional latent representation
Hidden units	8	Network capacity for feature extraction
Clusters	3	Number of regional risk groups
Pretraining epochs	100	Initial autoencoder training iterations
Clustering epochs	100	Fine-tuning iterations for clustering
Batch size	8	Samples per weight update
Optimizer	Adam (0.001)	Adaptive optimization with a stable learning rate

Loss (pretrain)	MSE	Input–output reconstruction loss
Loss (clustering)	KL Divergence	Optimization of cluster distribution
Initialization	K-Means	Initial cluster centroid assignment

In the experimental phase, the DEC model was configured using three input features representing key stunting indicators. Dimensionality reduction was performed through an autoencoder architecture consisting of eight hidden units and a two-dimensional latent space to generate compact feature representations.

The training process consisted of two stages: a pretraining phase conducted over 100 epochs using Mean Squared Error (MSE) to learn the intrinsic structure of the data, followed by a clustering phase of 100 epochs optimized using Kullback–Leibler (KL) divergence to refine cluster separation.

Model parameters were updated using the Adam optimizer with a learning rate of 0.001 and a batch size of 8 to ensure stable convergence. Cluster centroids were initialized using K-Means, with the number of clusters set to three based on prior evaluation results.

2.1 Hierarchical Clustering (HC)

Hierarchical Clustering (HC) constructs a hierarchical cluster structure based on pairwise distance measurements, resulting in a dendrogram that illustrates relationships among objects at different levels of similarity [15]. In this study, the agglomerative approach is employed, where each observation initially forms an individual cluster and is iteratively merged with other clusters according to a predefined linkage criterion [16]. The merging process continues until desired number of clusters is achieved [17].

This clustering concept requires the calculation of dissimilarity distances between data points. One of the most commonly used methods for measuring dissimilarity is Euclidean distance, as expressed in Equation 1 [18], [19]:

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

Agglomerative clustering works by grouping data in a bottom-up manner. Initially, each data point is treated as an individual cluster (leaf) consisting of a single member [20], [21]. To produce clusters with high internal consistency, Ward's linkage method is applied, which minimizes the increase in total within-cluster variance (ΔESS) at each merging stage, as expressed in Equation 2 [22]:

$$\Delta ESS = \frac{n_A n_B}{n_A + n_B} \|\bar{x}_A - \bar{x}_B\|^2 \quad (2)$$

Where n_A and n_B represent the number of members in clusters A and B, respectively, while $\bar{x}_A$dan $\bar{x}_B$ denote the average vectors of each cluster. The smallest ΔESS value indicates the most optimal merger.

The advantage of the Ward's method lies in its ability to generate clusters that tend to be balanced in size and compact in shape, making it particularly suitable for normalized numerical datasets [9].

2.2 Deep Embedded Clustering (DEC)

Deep Embedded Clustering (DEC) is a representation learning–based clustering method that integrates deep neural networks with clustering optimization. Unlike conventional clustering techniques that operate directly in the original feature space, DEC first transforms the input data into a lower-dimensional latent representation using an autoencoder architecture [23], [24], [21].

The model utilizes an autoencoder architecture comprising an encoder and a decoder. The encoder maps the input data (x) into a low-dimensional latent representation (z), which is subsequently used for initial clustering. The architecture employs fully connected layers with ReLU activation functions to accelerate training convergence and mitigate the vanishing gradient problem. The ReLU activation function is defined in Equation 3:

$$f(x) = \max(0, x) \quad (3)$$

The decoder reconstructs input data from the latent space in order to to minimize reconstruction error. Simultaneously, the model is optimized by minimizing the Kullback–Leibler (KL) divergence loss between the soft assignment distribution q_{ij}, representing the probability that data point i belongs to cluster j, and the target distribution p_{ij}. This optimization process refines cluster formation within the latent space, as formulated in Equation 4:

$$L = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}} \quad (4)$$

The target distribution P is constructed by emphasizing the contribution of data points with high confidence toward a particular cluster, thereby enabling the learning process to become more stable and representative.

2.3 Evaluation of Measured Outcomes

The clustering results were analyzed and evaluated to obtain a deeper understanding of the similarity patterns and characteristic relationships among regions. The evaluation was based on the Silhouette Score metric, which measures how well an object is assigned to its corresponding cluster [25], [26].

The Silhouette Score for each data point i is calculated using two components: $a(i)$, which represents the average distance between data point i and all other points within the same cluster, and $b(i)$, which represents the minimum average distance between data point i and points in the nearest neighboring cluster. The silhouette value for data point i is formulated in Equation 5:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \quad (5)$$

The value of $s(i)$ lies within the range [-1, 1], where values close to 1 indicate that the data point is appropriately assigned to its cluster, values close to 0 indicate that the data point lies near the boundary between clusters, and negative values indicate potential misclassification or overlap between clusters [27], [28].

3. Results and Discussion

3.1 Dataset

The data used in this study were obtained through a secondary data collection process originating from health agencies located in districts across Madura Island. The dataset includes information on stunting conditions from three regencies on Madura Island: Bangkalan Regency, Sampang Regency, and Pamekasan Regency, covering the period from 2020 to 2024.

Spatially, the dataset covers three regencies:

- Bangkalan Regency, consisting of 19 sub-districts (Kamal, Labang, Kwanyar, Kedundung, Modung, Blega, Konang, Galis, Tanah Merah, Tragah, Socah, Bangkalan, Burneh, Arosbaya, Geger, Kokop, Tanjung Bumi, Sepulu, and Klampis),
- Sampang Regency, consisting of 14 sub-districts (Sreseh, Torjun, Pangarengan, Sampang, Camplong, Omben, Kedundung, Jrengik, Tambelangan, Banyuates, Robatal, Karang Penang, Ketapang, and Sokobanah), and
- Pamekasan Regency, consisting of 13 sub-districts (Tlanakan, Pademawu, Galis, Larangan, Pamekasan, Propo, Palengan, Pegantengan, Kadur, Pakong, Waru, Batumarmar, and Pasean).

Data from each health agency were systematically integrated into a single unified dataset that served as the basis for the clustering analysis. Each data entry represents a subdistrict-level administrative area within its respective regency. Consequently, a total of 46 administrative units were used as the objects of analysis.

Each subdistrict contains annual data spanning five years (2020–2024) with six primary variables, as described in Table 2, while an excerpt of the dataset is presented in Table 3. The primary variables used in the analysis include the number of toddlers measured, the number of stunting cases, and the percentage of stunting prevalence. Meanwhile, the regency and subdistrict variables function as spatial identifiers within the clustering analysis.

Table 2. Dataset Structure

Variable	Description
Regency	Name of district in Madura (Bangkalan, Sampang, Pamekasan)
Subdistrict	Name of subdistrict as the unit of analysis
Number of Toddlers Measured	Number of toddlers whose nutritional status was measured
Number of Stunting	Number of toddlers identified as having stunting
Stunting Percentage	Percentage of stunted toddlers to the total number of toddlers measured
Years	Year of data collection (2020–2024)

Table 3. Data Fragment

Regency	Subdistrict	Number of Toddlers Measured	Number of Stunting	Stunting Percentage	Years
PAMEKASAN	TLANAKAN	4775	202	4,230366	2024
PAMEKASAN	PADEMAWU	5542	142	2,562251	2024

PAMEKASAN	GALIS	1999	29	1,450725	2024
PAMEKASAN	LARANGAN	4602	105	2,281616	2024
PAMEKASAN	PAMEKASAN	6159	295	4,789738	2024
…	…	…	…	…	…
BANGKALAN	GEGER	3093	67	2.2	2024
BANGKALAN	KOKOP	3747	155	4.1	2024
BANGKALAN	TANJUNG BUMI	2203	121	5.5	2024
BANGKALAN	SEPULU	2026	32	1.6	2024
BANGKALAN	KLAMPIS	2984	25	0.8	2024

3.2 Data Preparation

Prior to the analysis process, data cleaning and normalization procedure were performed to ensure that all numerical values were correctly interpreted. This stage included checking for missing values, converting data types from string to float, and standardizing decimal number formatting from commas to periods. Subsequently, scale normalization was applied to ensure all variables were represented within a comparable range. Following these preprocessing steps, EDA was conducted to examine the overall distribution.

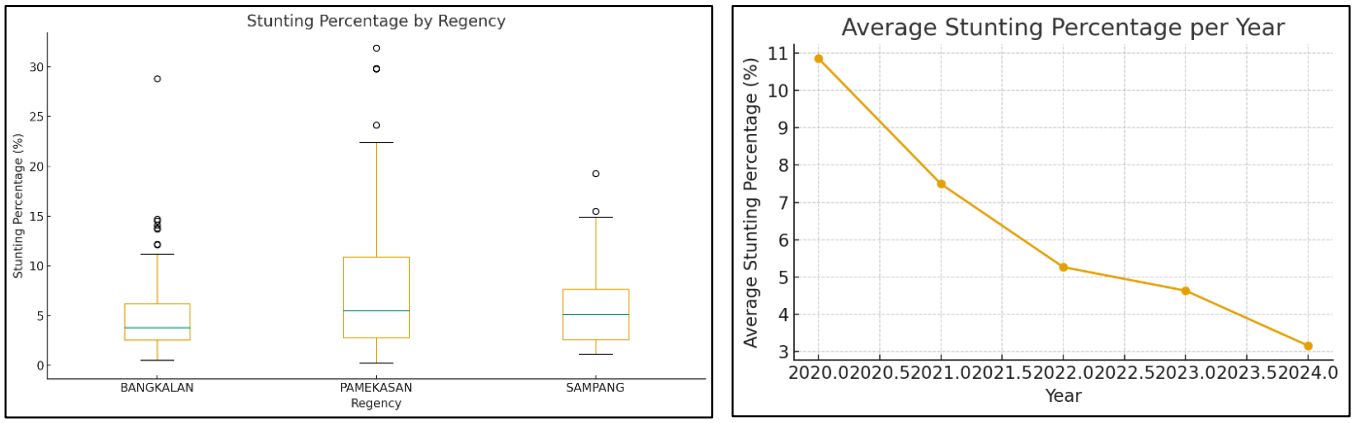
Based on the initial data exploration results, the boxplot visualization presented in Figure 2(a) shows significant differences in the distribution of stunting prevalence across regions. The value distribution shown in the boxplot indicates that several regions exhibit higher median stunting percentages, suggesting that most subdistricts within these regions continue to experience relatively severe stunting conditions. In contrast, other regions display narrower value ranges with lower median values, reflecting a more stable and balance condition in stunting management.

The presence of outliers above the upper whisker further indicates that several subdistricts experience extreme stunting prevalence levels exceeding 25–30%, requiring special attention in intervention programs.

Furthermore, the visualization of the average annual stunting prevalence presented in Figure 2(b) provides an overview of the temporal development of stunting conditions during the 2020-2024 period. The results show a gradual decline in average stunting prevalence over time, which may indicate that various nutrition and child health intervention programs have begun to produce positive outcomes.

Descriptively, the exploration results show that stunting prevalence across regions in Madura Island exhibits substantial variability. This variation indicates the existence of heterogeneous regional characteristics potentially influenced by socioeconomic conditions, availability of healthcare facilities, and community nutritional consumption patterns.

These variability patterns were then subsequently analyzed further using clustering methods to identify groups of regions with similar stunting risk characteristics.



(a) (b)

Figure 2. Data Distribution

3.3 Clustering Results

The HC method was used to identify interregional similarities based on the characteristics of stunting-related variables. This study employed an agglomerative clustering approach, where each region was initially treated as an individual cluster and subsequently merged progressively according to similarity measured using Euclidean distance.

The cluster merging criterion was determined using Ward's linkage method, which minimizes the total variance among cluster members. Initially, the optimal number of clusters (k) was determined to identify the most appropriate grouping configuration for the dataset. Table 4 presents the results of the optimal number of clusters for each year.

Table 4. Results of Determining the Optimal Number of Clusters in HC

Year	Silhouette Score of k			
	2	3	4	5
2020	0.4844	0.4851	0.4290	0.3343
2021	0.4985	0.5430	0.5249	0.4326
2022	0.4433	0.4059	0.4001	0.3890
2023	0.4466	0.4938	0.3845	0.3534
2024	0.4614	0.5314	0.4355	0.4704

Based on the results shown in Table 4, the three-cluster configuration ($k = 3$) produced the highest Silhouette Score in most years, particularly 2020 (0.4851), 2021 (0.5430), 2023 (0.4938), and 2024 (0.5314). These findings indicate that dividing the regions into three categories provides the most effective configuration for distinguishing stunting risk characteristics, namely low-risk, medium-risk, and high-risk clusters.

Meanwhile, in 2022, the optimal configuration consisted of two clusters representing low-risk and medium-risk categories. The clustering results are visualized using scatter plots (Figure 3) and dendrograms (Figure 4), which facilitate interpretation of the optimal cluster structures. Detailed compositions of the cluster formed are presented in Table 5.

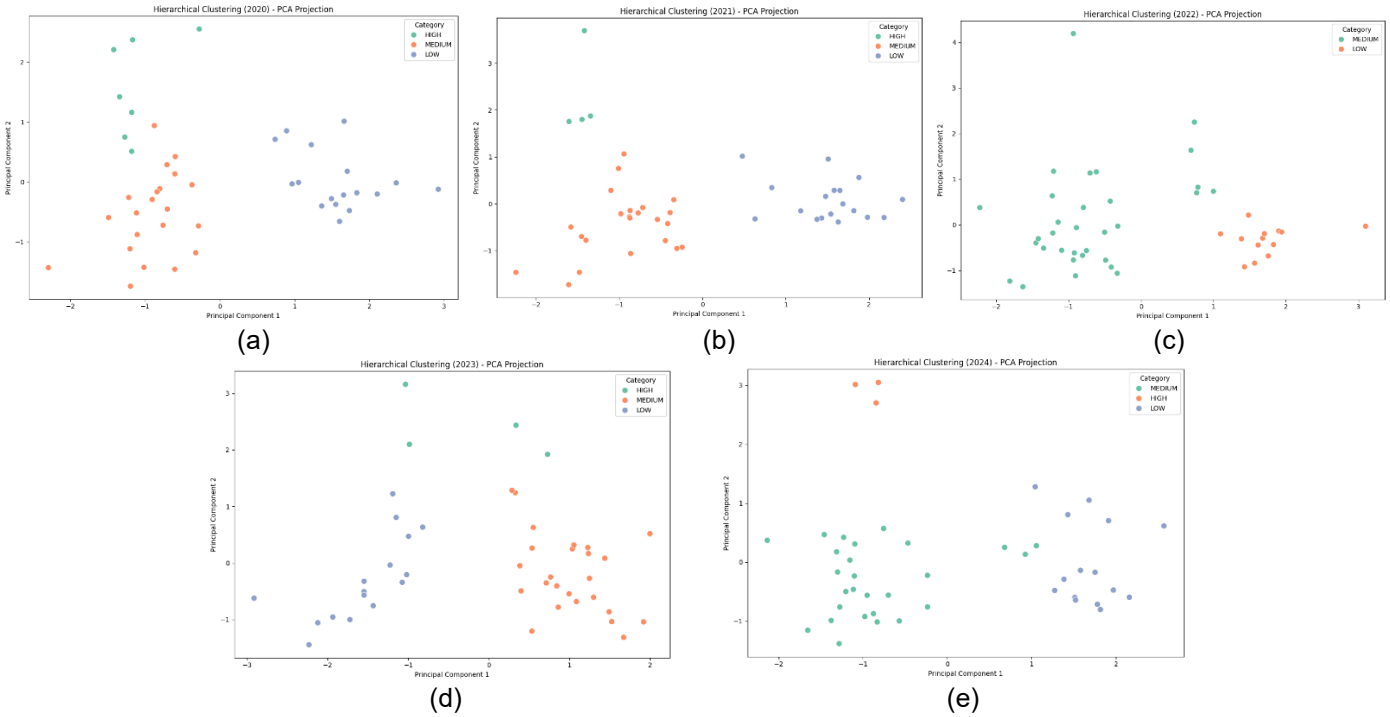
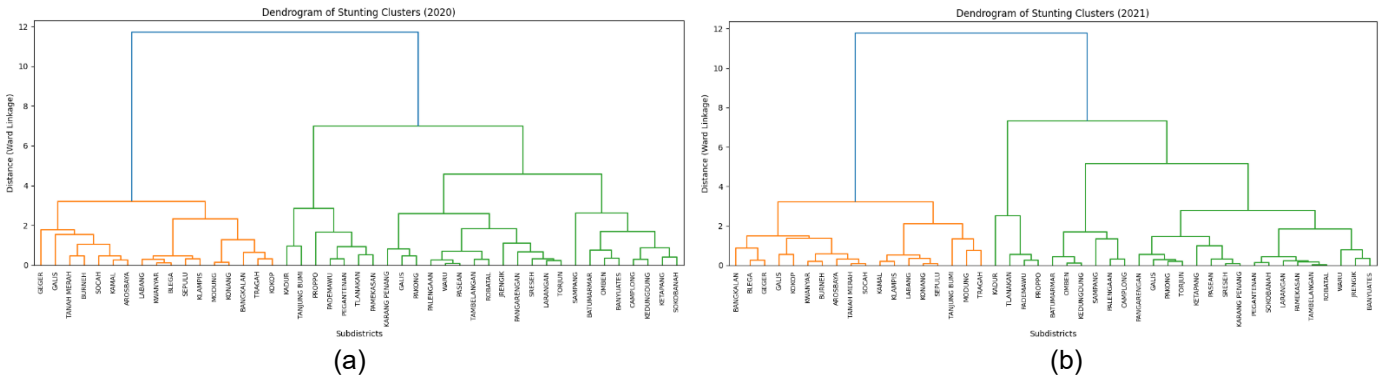


Figure 3. Cluster Identification Results Using HC



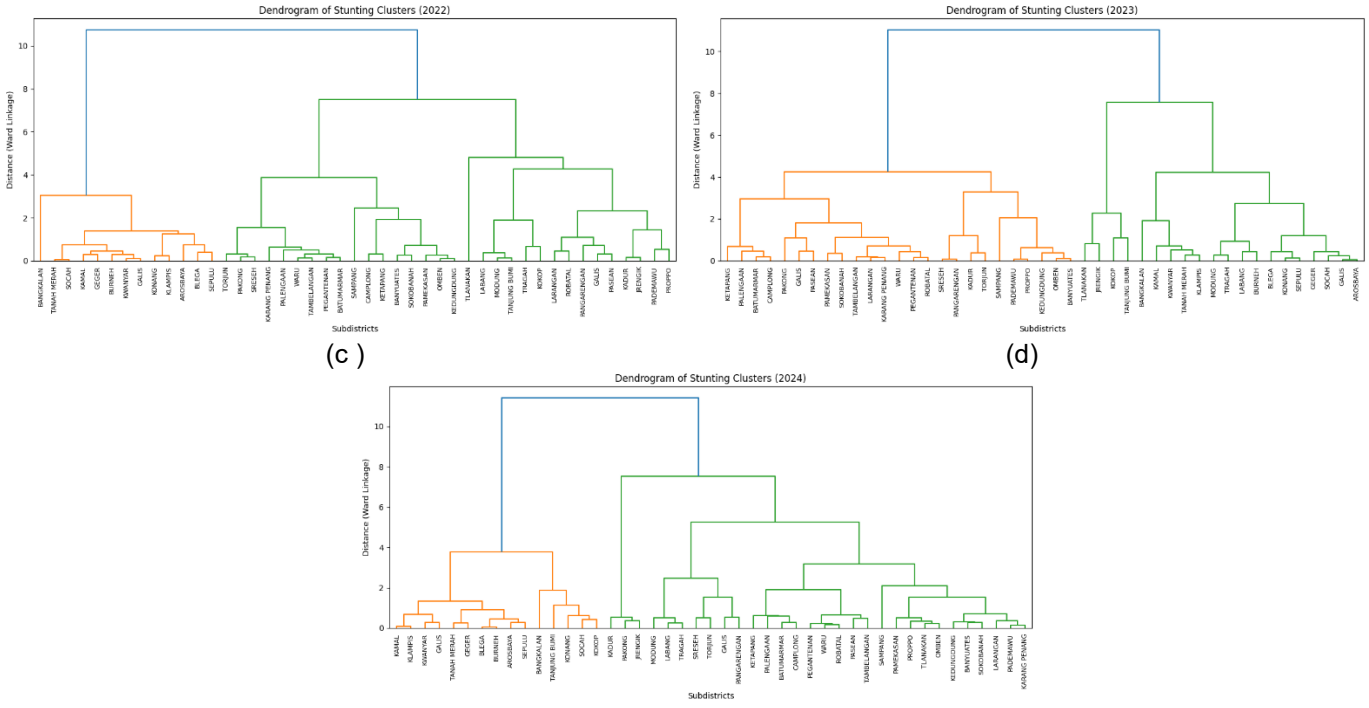


Figure 4. HC Result Dendrogram

Table 5. Detailed Cluster Composition Using HC

Year	Description	Cluster		
		Low	Medium	High
2020	Mean prevalence	7.26%	9.33%	24.44%
	Total subdistrict	17	21	7
	Subdistricts	KAMAL, LABANG, KWANYAR, MODUNG, BLEGA, KONANG, GALIS, TANAH MERAH, TRAGAH, SOCAH, BANGKALAN, BURNEH, AROSBAYA, GEGER, KOKOP, SEPULU, KLAMPIS	GALIS, LARANGAN, PALENGAAN, PAKONG, WARU, BATUMARMAR, PASEAN, SRESEH, TORJUN, PANGARENGAN, SAMPANG, CAMPLONG, OMBEN, KEDUNGDUNG, JRENGIK, TAMBELANGAN, BANYUATES, ROBATAL, KARANG PENANG, KETAPANG, SOKOBANAH	TLANAKAN, PADEMAWU, PAMEKASAN, PROPPO, PEGANTENAN, KADUR, TANJUNG BUMI

Year	Description	Cluster		
		Low	Medium	High
2021	Mean prevalence	4.78%	6.66%	24.16%
	Total subdistrict	18	23	4
	Subdistricts	KAMAL, LABANG, KWANYAR, MODUNG, BLEGA, KONANG, GALIS, TANAH MERAH, TRAGAH, SOCAH, BANGKALAN, BURNEH, AROSBAYA, GEGER, KOKOP, TANJUNG BUMI, SEPULU, KLAMPIS	GALIS, LARANGAN, PAMEKASAN, PALENGAAN, PEGANTENAN, PAKONG, WARU, BATUMARMAR, PASEAN, SRESEH, TORJUN, PANGARENGAN, SAMPANG, CAMPLONG, OMBEN, KEDUNGDUNG, JRENGIK, TAMBELANGAN, BANYUATES, ROBATAL, KARANG PENANG, KETAPANG, SOKOBANAH	TLANAKAN, PADEMAWU, PROPPO, KADUR
2022	Mean prevalence	3.00%	6.23%	-
	Total subdistrict	13	32	
	Subdistricts	KAMAL, KWANYAR, BLEGA, KONANG, GALIS, TANAH MERAH, SOCAH, BANGKALAN, BURNEH, AROSBAYA, GEGER, SEPULU, KLAMPIS	TLANAKAN, PADEMAWU, GALIS, LARANGAN, PAMEKASAN, PROPPO, PALENGAAN, PEGANTENAN, KADUR, PAKONG, WARU, BATUMARMAR, PASEAN, SRESEH, TORJUN, PANGARENGAN, SAMPANG, CAMPLONG, OMBEN, KEDUNGDUNG, JRENGIK, TAMBELANGAN, BANYUATES, ROBATAL, KARANG PENANG, KETAPANG, SOKOBANAH, LABANG, MODUNG, TRAGAH, KOKOP, TANJUNG BUMI	
2023	Mean prevalence	3.78%	3.96%	12.24%
	Total subdistrict	16	25	4
	Subdistricts	KAMAL, LABANG, KWANYAR, MODUNG, BLEGA, KONANG, GALIS, TANAH MERAH, TRAGAH, SOCAH, BANGKALAN, BURNEH,	PADEMAWU, GALIS, LARANGAN, PAMEKASAN, PROPPO, PALENGAAN, PEGANTENAN, KADUR, PAKONG,	TLANAKAN, JRENGIK, KOKOP, TANJUNG BUMI

Year	Description	Cluster		
		Low	Medium	High
		AROSBAYA, GEGER, SEPULU, KLAMPIS	WARU, BATUMARMAR, PASEAN, SRESEH, TORJUN, PANGARENGAN, SAMPANG, CAMPLONG, OMBEN, KEDUNGDUNG, TAMBELANGAN, BANYUATES, ROBATAL, KARANG PENANG, KETAPANG, SOKOBANAH	
2024	Mean prevalence	2.52%	2.80%	9.69%
	Total subdistrict	15	27	3
	Subdistricts	KAMAL, KWANYAR, BLEGA, KONANG, GALIS, TANAH MERAH, SOCAH, BANGKALAN, BURNEH, AROSBAYA, GEGER, KOKOP, TANJUNG BUMI, SEPULU, KLAMPIS	TLANAKAN, PADEMAWU, GALIS, LARANGAN, PAMEKASAN, PROPPO, PALENGAAN, PEGANTENAN, WARU, BATUMARMAR, PASEAN, SRESEH, TORJUN, PANGARENGAN, SAMPANG, CAMPLONG, OMBEN, KEDUNGDUNG, TAMBELANGAN, BANYUATES, ROBATAL, KARANG PENANG, KETAPANG, SOKOBANAH, LABANG, MODUNG, TRAGAH	KADUR, PAKONG, JRENGIK

DEC integrates a neural network-based autoencoder with the K-Means algorithm. This approach enables the learning of latent data representations, which are subsequently used for clustering within a lower-dimensional space. Evaluation of the clustering results using the Silhouette Score demonstrates performance variation across different years, as presented in Table 6, while the clustering distributions are visualized in the scatter plot shown in Figure 5.

The highest Silhouette Score was obtained in 2024 (0.4874), whereas the lowest score occurred in 2021 (0.1670). This trend indicates progressive improvement in clustering quality over time, suggesting that the DEC model became increasingly capable of differentiating regional characteristics based on stunting risk levels. The relatively low scores observed during the initial years (2020–2021) indicate high similarity patterns among regions, whereas the increasing scores in subsequent years suggest that DEC gradually succeeded in capturing nonlinear structures and latent patterns that were not detectable using conventional clustering methods.

Detailed compositions of each cluster generated by the DEC method are presented in Table 7.

Table 6. DEC Method Evaluation Results

	Year				
	2020	2021	2022	2023	2024
Silhouette Score	0,3144	0,1670	0,4573	0,4122	0,4874

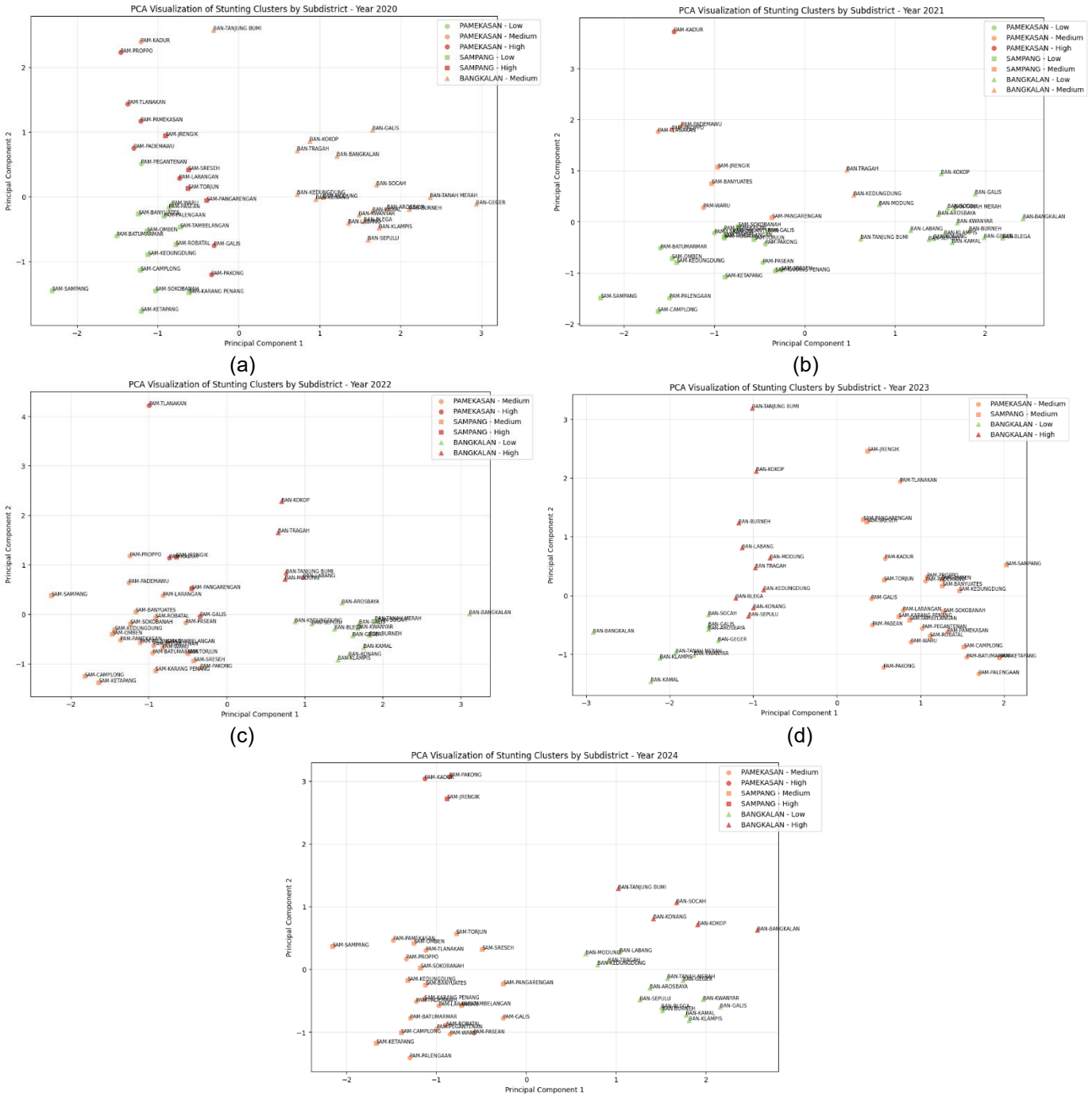


Figure 5. Cluster Identification Results Using DEC

Table 7. Detailed Cluster Composition Using DEC

Year	Description	Cluster		
		Low	Medium	High
2020	Mean prevalence	8.75%	9.56%	16.10%
	Total subdistrict	15	20	11
	Subdistricts	PALENGAAN, PEGANTENAN, WARU, BATUMARMAR, PASEAN, SAMPANG, CAMPLONG, OMBEN,	KADUR, KAMAL, LABANG, KWANYAR, KEDUNGDUNG, MODUNG, BLEGA, KONANG, GALIS,	TLANAKAN, PADEMAWU, GALIS, LARANGAN, PAMEKASAN, PROPPO, PAKONG,

Year	Description	Cluster		
		Low	Medium	High
		KEDUNGDUNG, TAMBELANGAN, BANYUATES, ROBATAL, KARANG PENANG, KETAPANG, SOKOBANAH	TANAH MERAH, TRAGAH, SOCAH, BANGKALAN, BURNEH, AROSBAYA, GEGER, KOKOP, TANJUNG BUMI, SEPULU, KLAMPIS	SRESEH, TORJUN, PANGARENGAN, JRENGIK
2021	Mean prevalence	4.95%	14.95%	31.89%
	Total subdistrict	36	9	1
	Subdistricts	GALIS, LARANGAN, PAMEKASAN, PALENGAAN, PEGANTENAN, PAKONG, BATUMARMAR, PASEAN, SRESEH, TORJUN, SAMPANG, CAMPLONG, OMBEN, KEDUNGDUNG, TAMBELANGAN, ROBATAL, KARANG PENANG, KETAPANG, SOKOBANAH, KAMAL, LABANG, KWANYAR, MODUNG, BLEGA, KONANG, GALIS, TANAH MERAH, SOCAH, BANGKALAN, BURNEH, AROSBAYA, GEGER, KOKOP, TANJUNG BUMI, SEPULU, KLAMPIS	TLANAKAN, PADEMAWU, PROPPO, WARU., PANGARENGAN, JRENGIK, BANYUATES, KEDUNGDUNG, TRAGAH	KADUR
2022	Mean prevalence	3.06%	4.40%	10.25%
	Total subdistrict	14	22	10
	Subdistricts	KAMAL, KWANYAR, KEDUNGDUNG, BLEGA, KONANG, GALIS, TANAH MERAH, SOCAH, BANGKALAN, BURNEH, AROSBAYA, GEGER, SEPULU, KLAMPIS	PADEMAWU, LARANGAN, PAMEKASAN, PROPPO, PALENGAAN, PEGANTENAN, PAKONG, WARU, BATUMARMAR, PASEAN, SRESEH, TORJUN, SAMPANG, CAMPLONG, OMBEN, KEDUNGDUNG, TAMBELANGAN, BANYUATES, ROBATAL, KARANG PENANG, KETAPANG, SOKOBANAH	TLANAKAN, GALIS, KADUR, PANGARENGAN, JRENGIK, LABANG, MODUNG, TRAGAH, KOKOP, TANJUNG BUMI
2023	Mean prevalence	2.37%	4.52%	6.97%
	Total subdistrict	9	27	10
	Subdistricts	KAMAL, KWANYAR, GALIS, TANAH MERAH, SOCAH, BANGKALAN,	TLANAKAN, PADEMAWU, GALIS, LARANGAN,	ABANG, KEDUNGDUNG, MODUNG, BLEGA,

Year	Description	Cluster		
		Low	Medium	High
		AROSBAYA, GEGER, KLAMPIS	PAMEKASAN, PROPPO, PALENGAAN, PEGANTENAN, KADUR, PAKONG, WARU, BATUMARMAR, PASEAN, SRESEH, TORJUN, PANGARENGAN, SAMPANG, CAMPLONG, OMBEN, KEDUNGDUNG, JRENGIK, TAMBELANGAN, BANYUATES, ROBATAL, KARANG PENANG, KETAPANG, SOKOBANAH	KONANG, TRAGAH, BURNEH, KOKOP, TANJUNG BUMI, SEPULU
2024	Mean prevalence Total subdistrict Subdistricts	1.97% 14 KAMAL, LABANG, KWANYAR, KEDUNGDUNG, MODUNG, BLEGA, GALIS, TANAH MERAH, TRAGAH, BURNEH, AROSBAYA, GEGER, SEPULU, KLAMPIS	2.75% 24 TLANAKAN, PADEMAWU, GALIS, LARANGAN, PAMEKASAN, PROPPO, PALENGAAN, PEGANTENAN, WARU, BATUMARMAR, PASEAN, SRESEH, TORJUN, PANGARENGAN, SAMPANG, CAMPLONG, OMBEN, KEDUNGDUNG, TAMBELANGAN, BANYUATES, ROBATAL, KARANG PENANG, KETAPANG, SOKOBANAH	6.47% 8 KADUR, PAKONG, JRENGIK, KONANG, SOCAH, BANGKALAN, KOKOP, TANJUNG BUMI

3.4 Comparative Analysis of HC and DEC

The comparison between HC and DEC was conducted using the Silhouette Score metric, which measures how well the data points are grouped within clusters and how distinct each cluster is from neighboring clusters (Figure 6). Overall, the results show that HC achieved more consistent and higher average Silhouette Score, ranging from 0.48 to 0.54, compared to DEC, which exhibited more variable scores ranging from 0.16 to 0.49. HC achieved its best performance in 2021 with a Silhoutte Score of 0.5430, whereas DEC reached its highest score in 2024 with a value of 0.4874.

These findings indicate that HC provides stronger clustering stability, while DEC demonstrates progressive improvement over time, reflecting its capability to adapt and learn more representative feature structures from increasingly complex data.

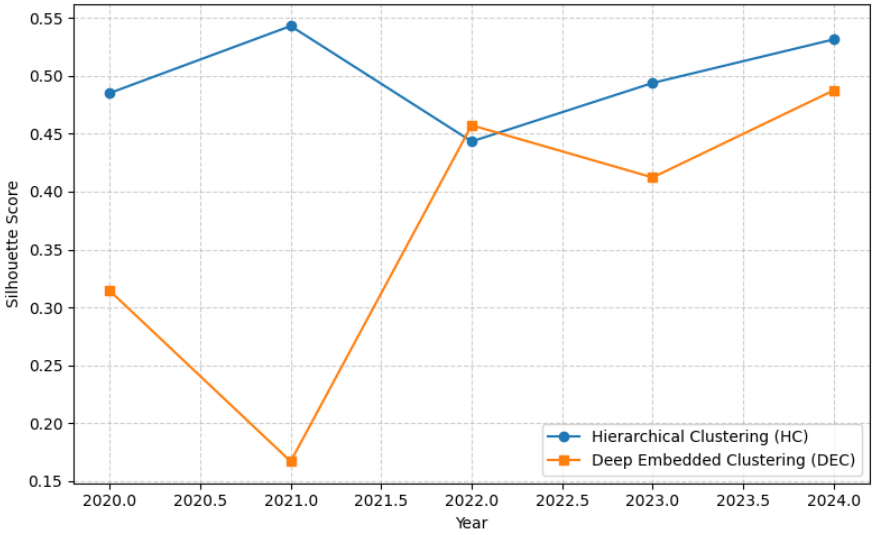


Figure 6. Comparison of Silhouette Score Values between HC and DEC

The comparison between HC and DEC was not limited to Silhouette Score values alone but also examined from the perspective of cluster stability, consistency of risk categorization, and policy relevance derived from the clustering outcomes. Quantitatively, HC demonstrated more stable performance, with relatively consistent Silhouette Scores that tended to remain higher across most observation years. This stability suggests that the structure of regional stunting risk during the study period can be effectively explained using similarity patterns based on Euclidean distance and linear variation among variables.

In other words, when the distribution of stunting prevalence across regions still exhibits relatively homogeneous patterns or clear proportional differences, the hierarchical clustering approach is sufficient for constructing regional risk stratification maps.

In contrast, DEC exhibited more dynamic performance variation across years. During the initial observation period, lower Silhouette Scores indicated that the nonlinear data structure had not yet been sufficiently separated or that latent representations had not provided substantial advantages over conventional clustering methods. However, the increasing scores observed in subsequent years suggest that DEC gradually succeeded in capturing more complex patterns, including potential latent interactions among the number of toddlers measured, the number of stunting cases, and stunting prevalence percentage. These findings indicates that when regional heterogeneity increases or when differences among regions are no longer entirely linear, a representation learning-based approach can provide more adaptive risk mapping.

From a policy perspective, the two methods offer distinct strategic advantages. HC is more suitable as a baseline regional risk mapping tool due to its stability and interpretability. The dendrogram structure and cluster consistency facilitate transparent identification of low-, medium-, and high-risk areas, making HCparticularly relevant for annual budget allocation planning, prioritization of nutritional interventions, and district-level public health monitoring.

Conversely, DEC demonstrates greater potential as a regional pattern-shift detection mechanism. Its ability to learn latent representations enables the identification of regions experiencing changes in risk structures that may not be fully observable using simple distance-based approaches. In the context of public health policymaking, this capability is important for detecting regions that begin to show increasing vulnerability even when absolute stunting prevalence has not yet reached critical levels.

Therefore, DEC can function as a data-driven early warning system to support more targeted preventive interventions. By utilizing insights generated from both clustering approach, policymakers can implement a dual-intervention strategy: prioritizing high-risk HC clusters for direct nutritional and healthcare support, while leveraging DEC-based early warnings to initiate preventive measures such as sanitation improvement programs and nutritional education in emerging high-risk areas.

This targeted data-driven framework enables more precise and evidence-based public health resource allocation, thereby improving the effectiveness of intervention policies and supporting early prevention of stunting escalation.

4. Conclusion

This study successfully identified patterns of stunting-prone regions on Madura Island using a comparative approach involving Hierarchical Clustering (HC) and Deep Embedded Clustering (DEC). Based on the analysis results,

Hierarchical Clustering demonstrated more stable and consistent performance throughout the 2020–2024 period, achieving Silhoutte Scores ranging from 0.40 to 0.54, which indicate moderate to good cluster quality.

These findings reinforce the assumption that a relatively consistent spatial pattern exists in the distribution of stunting-prone regions, where several areas exhibit similar stunting characteristics across multiple years. Meanwhile, the Deep Embedded Clustering method showed an increasing performance trend over time, achieving its highest Silhouette Score of 0.4874 in 2024. This result indicates DEC's capability to learn more complex nonlinear representations and capture spatial-temporal variations within stunting data.

Although its initial performance was lower than that of HC, DEC shows strong potential for application in large-scale and heterogeneous stunting data analysis, particularly when the data structure cannot be adequately represented using conventional distance-based relationships.

Acknowledgement

The author would like to express their sincere gratitude to University of Trunojoyo Madura for providing the opportunity and support to conduct this research. The authors also thank the Central Statistics Agency and Health offices in the Madura Island region for their assistance in providing data used in this research.

References

[1] Health Development Policy Agency of the Ministry of Health of the Republic of Indonesia, "Data Catalog: Indonesian Nutrition Status Survey (SSGI) 2022," Indonesia, 2022.
[2] Health Development Policy Agency of the Ministry of Health of the Republic of Indonesia, "Data Catalog: Indonesian Health Survei," Indonesia, 2023.
[3] Acceleration of Stunting Prevention/TP2AK, "Baseline Report of the 2018-2024 Stunting Prevention Acceleration Program," Indonesia, 2021.
[4] R. Mishra and S. Bera, "Geospatial and environmental determinants of stunting, wasting, and underweight: Empirical evidence from rural South and Southeast Asia," *Nutrition*, vol. 120, p. 112346, 2024. https://doi.org/10.1016/j.nut.2023.112346
[5] S. A. Bhat and N.-F. Huang, "Big data and ai revolution in precision agriculture: Survey and challenges," *Ieee Access*, vol. 9, pp. 110209–110222, 2021. https://doi.org/10.1109/ACCESS.2021.3102227
[6] Y. Shi, "Advances in big data analytics," *Adv Big Data Anal*, vol. 10, pp. 978–981, 2022. https://doi.org/10.1007/978-981-16-3607-3
[7] H. B. Abdalla, "A brief survey on big data: technologies, terminologies and data-intensive applications," *J. Big Data*, vol. 9, no. 1, p. 107, 2022. https://doi.org/10.1186/s40537-022-00659-3
[8] T. T. Khoei and A. Singh, "Data reduction in big data: a survey of methods, challenges and future directions," *Int. J. Data Sci. Anal.*, vol. 20, no. 3, pp. 1643–1682, 2025. https://doi.org/10.1007/s41060-024-00603-z
[9] J. Han, M. Kamber, and J. Pei, "Data mining: Concepts and," *Techniques, Waltham: Morgan Kaufmann Publishers*, 2012.
[10] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," *Artif. Intell. Rev.*, vol. 56, no. 8, pp. 8219–8264, 2023. https://doi.org/10.1007/s10462-022-10366-3
[11] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*, PMLR, 2016, pp. 478–487. https://doi.org/10.48550/arXiv.1511.06335
[12] F. E. Harrell and D. G. Levy, "Regression modeling strategies," *R package version*, pp. 3–6, 2022. https://doi.org/10.1007/978-3-319-19425-7
[13] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "Cluster analysis," 2011.
[14] E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long, "A survey of clustering with deep learning: From the perspective of network architecture," *IEEE access*, vol. 6, pp. 39501–39514, 2018. https://doi.org/10.1109/ACCESS.2018.2855437
[15] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Education India, 2016.
[16] A. Annisa, Y. Munarko, and Y. Azhar, "Peringkasan Tweet Berdasarkan Trending Topic Twitter Dengan Pembobotan TF-IDF dan Single Linkage Angglomerative Hierarchical Clustering," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, pp. 9–16, 2016. https://doi.org/10.22219/kinetik.v1i1.7
[17] O. Maimon and L. Rokach, *Data mining and knowledge discovery handbook*, vol. 2, no. 2005. Springer, 2005. https://doi.org/10.1007/b107408
[18] F. Damayanti, S. Herawati, I. Imamah, and A. Rachmad, "Indonesian license plate recognition based on area feature extraction," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 17, no. 2, pp. 620–627, 2019. http://doi.org/10.12928/telkomnika.v17i2.9017
[19] F. A. Mufarroha and F. Utaminingrum, "Hand gesture recognition using adaptive network based fuzzy inference system and K-nearest neighbor," *International Journal of Technology*, vol. 8, no. 3, pp. 559–567, 2017. https://doi.org/10.14716/ijtech.v8i3.3146
[20] R. T. Adek, R. K. Dinata, and A. Ditha, "Online newspaper clustering in Aceh using the agglomerative hierarchical clustering method," *International Journal of Engineering, Science and Information Technology*, vol. 2, no. 1, pp. 70–75, 2022. https://doi.org/10.52088/ijesty.v2i1.206
[21] I. Shafi *et al.*, "A review of approaches for rapid data clustering: Challenges, opportunities, and future directions," *IEEE Access*, vol. 12, pp. 138086–138120, 2024. https://doi.org/10.1109/ACCESS.2024.3461798
[22] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *J. Am. Stat. Assoc.*, vol. 58, no. 301, pp. 236–244, 1963.
[23] H. Hadipour, C. Liu, R. Davis, S. T. Cardona, and P. Hu, "Deep clustering of small molecules at large-scale via variational autoencoder embedding and K-means," *BMC Bioinformatics*, vol. 23, no. Suppl 4, p. 132, 2022. https://doi.org/10.1186/s12859-022-04667-1
[24] M. Li, C. Cao, C. Li, and S. Yang, "Deep embedding clustering based on residual autoencoder," *Neural Process. Lett.*, vol. 56, no. 2, p. 127, 2024. https://doi.org/10.1007/s11063-024-11586-0
[25] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987. https://doi.org/10.1016/0377-0427(87)90125-7
[26] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, p. 759, 2021. https://doi.org/10.3390/e23060759
[27] H.-H. Tan, Y.-F. Tan, W.-H. Tan, and C.-P. Ooi, "Investigating Data Consistency in the ASHRAE Dataset Using Clustering and Label Matching," *IEEE Access*, 2025. https://doi.org/10.1109/ACCESS.2025.3615311
[28] S. Alrabie and A. Barnawi, "Enhancing Heart Sound Classification with Iterative Clustering and Silhouette Analysis: An Effective Preprocessing Selective Method to Diagnose Rare and Difficult Cardiovascular Cases," *Computer Modeling in Engineering & Sciences*, vol. 144, no. 2, p. 2481, 2025. https://doi.org/10.32604/cmes.2025.067977