An Improved Hybrid CURE–SNE Model for High-Dimensional Data Clustering

Dewi Sartika Br Ginting, T.H.F. Harumy, Ade Sarah Huzaifah, Ivanny Putri Marianto
Department of Computer Science and Technology Information, Universitas Sumatera Utara, Medan, Indonesia

***Abstract*—Stunting remains a critical public health issue in rural communities, largely driven by inadequate nutrition, poor sanitation, and unfavorable socioeconomic conditions. This study proposes a hybrid clustering approach by integrating Clustering Using Representatives (CURE) with t-distributed Stochastic Neighbor Embedding (t-SNE) to analyze stunting prevalence and support the optimization of child nutrition strategies. Secondary data were collected from publicly accessible national health and nutrition repositories, comprising 500 child records with multiple parameters, including anthropometric indicators, nutritional intake, maternal characteristics, environmental sanitation, and socioeconomic factors. The t-SNE algorithm was employed to reduce the high-dimensional data into a two-dimensional space while preserving neighborhood structures, followed by the application of the CURE algorithm to construct clusters that are robust to noise and outliers. Experimental results indicate that the proposed CURE–SNE approach successfully formed four distinct clusters, namely C1 Very High Stunting Risk with 128 data points (25.6%), C2 High Stunting Risk with 142 data points (28.4%), C3 Moderate/Transitional Stunting Risk with 117 data points (23.4%), and C4 Low Stunting Risk with 113 data points (22.6%). Cluster quality evaluation demonstrates that the hybrid CURE–SNE method achieves a higher Silhouette Score and a lower Davies Bouldin Index compared to the CURE only approach, indicating improved cluster separation and compactness. These findings confirm that combining dimensionality reduction with representative-based clustering enhances the interpretability of stunting patterns and provides a reliable analytical foundation for designing targeted and data-driven child nutrition interventions in rural settings.**

***Keywords*—Hybrid clustering; CURE-SNE; stunting; davies bouldin index; silhouete score**

I. Introduction

Stunting remains one of the most persistent and complex public health challenges affecting child development, particularly in rural communities of developing and low- to middle-income countries [24]. Defined as impaired growth and development resulting from chronic malnutrition, repeated infections, and inadequate psychosocial stimulation during early childhood, stunting reflects long-term nutritional deprivation rather than short-term dietary insufficiency [26]. The consequences of stunting extend beyond physical growth limitations, encompassing cognitive impairment, reduced educational achievement, increased susceptibility to disease, and diminished economic productivity in adulthood [28]. As a result, stunting not only represents an individual health issue but also poses a significant barrier to sustainable social and economic development.

Despite extensive efforts by governments and international organizations, the prevalence of stunting remains high in many rural areas due to a combination of structural, socioeconomic, and environmental factors [25]. Limited access to healthcare services, poor sanitation, low parental education, food insecurity, and poverty contribute to the uneven distribution of stunting cases across regions. These conditions create complex and heterogeneous patterns of malnutrition that vary significantly between communities [29]. Consequently, uniform intervention strategies often fail to address local needs effectively, highlighting the importance of targeted and data-driven approaches for designing nutrition education and intervention programs [30] [31]. In recent years, the increasing availability of large-scale health, demographic, and nutritional datasets has opened new opportunities for applying advanced data analytics and artificial intelligence techniques in public health research [23]. Data-driven methods allow researchers and policymakers to analyze multidimensional data and identify latent patterns that are difficult to capture using conventional statistical techniques [2]. Among these methods, unsupervised learning, particularly clustering, has emerged as a powerful approach for exploring stunting prevalence by grouping regions or populations with similar characteristics [16]. Clustering analysis can support evidence-based decision-making by enabling the identification of high-risk areas, prioritization of interventions, and optimization of resource allocation.

However, the application of traditional clustering algorithms to real-world public health data presents several challenges [21]. Algorithms such as K-Means are sensitive to outliers, require prior knowledge of the number of clusters, and assume spherical cluster shapes, which may not align with the inherent structure of health and nutrition datasets [8]. Hierarchical clustering methods, while more flexible, often suffer from high computational complexity and limited scalability when applied to large datasets [22]. Furthermore, stunting-related data are typically high-dimensional, noisy, and non-linearly distributed, making effective cluster formation and interpretation particularly difficult.

To overcome these limitations, more robust and adaptive clustering approaches are required. Clustering Using Representatives (CURE) is a hierarchical clustering algorithm specifically designed to handle large datasets, arbitrarily shaped clusters, and outliers by representing each cluster using multiple well-scattered representative points [9]. Unlike traditional hierarchical methods, CURE reduces sensitivity to noise and provides improved cluster stability [3]. Nevertheless,

when applied directly to high-dimensional data, CURE may still face challenges related to computational efficiency and cluster visualization [6]. Dimensionality reduction techniques offer a promising solution to these issues [10]. Among them, t-distributed Stochastic Neighbor Embedding (t-SNE) has gained considerable attention for its ability to project high-dimensional data into a low-dimensional space while preserving local neighborhood relationships [14]. t-SNE is particularly effective for visualizing complex, non-linear structures and revealing intrinsic data patterns [15]. However, t-SNE alone is not a clustering algorithm; rather, it serves as a preprocessing or visualization tool that can enhance the performance of clustering methods when used in combination [17]. In this study, a hybrid clustering framework that integrates t-SNE with the CURE algorithm is proposed to analyze stunting prevalence in rural communities [18]. The t-SNE technique is first employed to reduce the dimensionality of stunting-related indicators, thereby improving data representation and computational efficiency [12]. Subsequently, the CURE algorithm is applied to the transformed data to identify meaningful clusters that reflect similarities in nutritional status and associated socioeconomic factors [20]. This hybrid CURE–t-SNE approach aims to improve cluster quality, robustness, and interpretability compared to conventional clustering methods [1]. The motivation for adopting a hybrid CURE–t-SNE framework lies in its ability to address the specific characteristics of stunting datasets, which often involve heterogeneous distributions, outliers, and complex interdependencies among variables. By combining the strengths of both methods, the proposed approach provides a more reliable analytical framework for uncovering hidden patterns in stunting prevalence data. The resulting clusters offer valuable insights into the spatial and demographic distribution of malnutrition, enabling the identification of distinct community profiles that require different intervention strategies.

This research focuses on the use of secondary data obtained from publicly available health and nutrition databases to ensure reproducibility and scalability [29]. The clustering results are evaluated using established internal validation metrics and compared with traditional clustering algorithms to demonstrate the effectiveness of the proposed method. Beyond methodological contributions, this study emphasizes the practical implications of clustering analysis for optimizing child nutrition strategies [27]. By translating cluster characteristics into actionable insights, the findings can support targeted nutrition education programs, community-based interventions, and evidence-informed policymaking in rural settings.

The primary contributions of this study are threefold. First, it introduces a hybrid CURE–t-SNE clustering framework tailored to the analysis of stunting prevalence data [19]. Second, it provides a comprehensive evaluation and comparison of the proposed method against widely used clustering techniques. Third, it demonstrates the relevance of advanced unsupervised learning methods in supporting data-driven nutrition strategies and sustainable public health interventions [33]. Ultimately, this research contributes to the growing body of literature on the application of intelligent computing techniques for addressing complex societal challenges, particularly in the context of child nutrition and rural development.

II. Related Works

Research on stunting and child malnutrition has traditionally relied on statistical analysis and epidemiological studies to identify key determinants and assess intervention outcomes [28]. Early studies primarily focused on descriptive statistics and regression-based models to examine the relationship between stunting prevalence and factors such as socioeconomic status, maternal education, sanitation, and food security [26]. While these approaches provide valuable insights into causal relationships, they often assume linear associations and struggle to capture complex interactions among multiple variables in heterogeneous populations, particularly in rural settings. With the increasing availability of large-scale health and demographic datasets, machine learning techniques have been increasingly adopted to enhance the analysis of nutritional status and child growth indicators [5]. Supervised learning methods, including decision trees, support vector machines, and neural networks, have been used to predict stunting risk based on demographic and nutritional variables. Although these models achieve relatively high predictive accuracy, they require labeled data and predefined outcome classes, which may not always be available or reliable in public health datasets. Moreover, supervised models are less suitable for exploratory analysis when the objective is to discover hidden patterns or groupings within the data.

Unsupervised learning, particularly clustering, has therefore gained attention as an effective approach for analyzing stunting prevalence and nutritional patterns without requiring labeled data [13]. Several studies have employed K-Means clustering to group regions or households based on anthropometric indicators and socioeconomic variables [7]. These studies demonstrate the potential of clustering to identify high-risk areas and support targeted interventions. However, K-Means assumes spherical clusters, is sensitive to initial centroids, and requires the number of clusters to be specified in advance. These limitations often result in unstable clusters when applied to complex and noisy health datasets [10]. Hierarchical clustering methods have also been widely used in nutritional and public health research due to their ability to produce a dendrogram that represents nested data structures. Agglomerative hierarchical clustering, in particular, has been applied to classify regions based on stunting prevalence and related indicators. While hierarchical methods offer greater flexibility in cluster shape compared to K-Means, they suffer from high computational complexity and limited scalability for large datasets. Additionally, the merging process in hierarchical clustering is irreversible, making the final cluster structure sensitive to early-stage decisions. To address the shortcomings of traditional clustering algorithms, density-based methods such as DBSCAN have been explored in some health-related studies. DBSCAN is capable of identifying arbitrarily shaped clusters and handling noise, making it suitable for datasets with irregular distributions. Nevertheless, its performance is highly dependent on parameter selection, and it may struggle with datasets that exhibit varying density

levels, which are common in stunting and nutritional data across diverse rural communities.

Clustering Using Representatives (CURE) was introduced as a robust hierarchical clustering algorithm designed to handle large datasets, outliers, and clusters with arbitrary shapes [32]. CURE represents each cluster using multiple representative points that are shrunk toward the cluster centroid, thereby reducing sensitivity to noise and extreme values [5]. Although CURE has been successfully applied in domains such as spatial data analysis and market segmentation, its application in public health and nutritional studies remains limited. This gap highlights the potential of CURE as an alternative clustering approach for stunting prevalence analysis, particularly in heterogeneous rural contexts. Another challenge in clustering stunting-related data lies in the high dimensionality of the datasets, which often include numerous nutritional, demographic, and socioeconomic variables [4]. High-dimensional data can degrade clustering performance due to the curse of dimensionality and reduced distance discrimination [11]. Dimensionality reduction techniques have therefore been increasingly incorporated into clustering frameworks. Principal Component Analysis (PCA) has been widely used to reduce dimensionality in health datasets; however, PCA primarily captures linear relationships and may fail to preserve complex non-linear structures inherent in nutritional data [13]. In contrast, t-distributed Stochastic Neighbor Embedding (t-SNE) has emerged as a powerful non-linear dimensionality reduction technique capable of preserving local neighborhood structures in low-dimensional space [14]. t-SNE has been extensively used for data visualization and exploratory analysis in biomedical and social science research. Several studies have demonstrated that integrating t-SNE with clustering algorithms can enhance cluster separation and interpretability. Nonetheless, most existing works utilize t-SNE in combination with K-Means or DBSCAN, with limited exploration of its integration with robust hierarchical algorithms such as CURE.

Recent hybrid clustering frameworks combining dimensionality reduction and clustering have shown promising results in complex data analysis [32]. These studies suggest that applying dimensionality reduction as a preprocessing step can improve clustering accuracy, reduce computational cost, and enhance visualization. However, in the context of stunting prevalence analysis, existing hybrid approaches remain underexplored and often lack comprehensive evaluation and comparison across multiple clustering algorithms. Based on the reviewed literature, a clear research gap can be identified. While clustering techniques have been applied to analyze stunting prevalence, most studies rely on conventional algorithms that struggle with high-dimensional, noisy, and heterogeneous data. The potential of robust hierarchical clustering methods such as CURE, particularly when combined with advanced non-linear dimensionality reduction techniques like t-SNE, has not been sufficiently investigated in the context of child nutrition and rural health analysis. This study addresses this gap by proposing and evaluating a hybrid CURE–SNE framework to improve clustering quality and support data-driven optimization of child nutrition strategies in rural communities.

III. Methodology

This study adopts a quantitative and experimental research methodology based on unsupervised machine learning to analyze stunting prevalence patterns in rural communities. A hybrid clustering framework integrating t-distributed Stochastic Neighbor Embedding (t-SNE) and Clustering Using Representatives (CURE) is proposed to address the challenges of high dimensionality, noise, and non-linear data distribution commonly found in nutritional and public health datasets. The overall methodology consists of data collection, preprocessing, dimensionality reduction, clustering, and evaluation stages.

A. Research Framework

The proposed methodology follows a sequential analytical pipeline designed to improve clustering quality and interpretability. Initially, secondary stunting-related data are collected from publicly available health databases. After preprocessing and normalization, t-SNE is applied to project the high-dimensional data into a lower-dimensional space. Subsequently, the CURE algorithm is employed to identify meaningful clusters. Finally, clustering performance is evaluated and compared with conventional clustering algorithms.

B. Dataset and Preprocessing

Let the original dataset be denoted as:

$$X_i = \{x_1, x_2, \ldots, x_n\}, X_i \in R^d$$

where,

- n represents the number of observations (regions or communities),
- d denotes the number of features (nutritional and socioeconomic indicators).

Before analysis, missing values are handled using mean imputation, and all features are normalized using Min–Max normalization to ensure equal contribution of each variable:

$$X'_{ij} = \frac{\mathrm{x_{ij}} - \min(x_j)}{\max(x_j) - min(x_j)}$$

C. Dimensionality Reduction Using t-SNE

t-distributed Stochastic Neighbor Embedding (t-SNE) is employed to reduce the dimensionality of the dataset while preserving local neighborhood structures. The similarity between two data points x_i and x_j In the high-dimensional space, it is defined as a conditional probability:

$$P_{j|i} = \frac{\exp\left(-\frac{\left||x_i - x_j|\right|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(\frac{\left||x_i - x_k|\right|^2}{2\sigma_i^2}\right)}$$

The joint probability is then computed as:

$$P_{ij} = \frac{P_{j|i} + P_{i|j}}{2n}$$

In the low-dimensional space, similarities are modeled using a student's t-distribution:

$$q_{ij} = \frac{\left(1 + \left||y_i - y_j\right|\right|^2\right)^2}{\sum_{k \neq l}\left(1 + \left||y_k - y_l\right|\right|^2\right)^{-1}}$$

The optimal low-dimensional embedding is obtained by minimizing the Kullback–Leibler (KL) divergence between the two distributions:

$$KL(P||Q) = \sum_{i \neq j} P_{ij} \log\left(\frac{P_{ij}}{Q_{ij}}\right)$$

The output of t-SNE is a reduced dataset:

$$Y = \{y_1, y_2, \ldots, y_n\}, y_i \in \mathbb{R}^k, k \ll d$$

t-SNE in this study is motivated by its ability to preserve local neighborhood structures in high-dimensional data, which is crucial for identifying complex patterns in stunting-related indicators. t-SNE is capable of capturing non-linear relationships among variables, making it well-suited as a preprocessing step before clustering. In this study, parameter tuning was performed on t-SNE, including the adjustment of perplexity, learning rate, and the number of iterations, in order to obtain an optimal data representation. The parameter selection was conducted empirically based on clustering quality evaluation. Furthermore, to ensure the robustness of the results, multiple experimental runs were carried out, and the findings demonstrated consistent clustering performance, as indicated by relatively stable Silhouette Score and Davies Bouldin Index (DBI) values across experiments.

D. Clustering Using Representatives (CURE)

The CURE algorithm is applied to the reduced dataset Yto identify clusters. Unlike traditional hierarchical clustering, CURE represents each cluster using a fixed number of well-scattered representative points. Initially, each data point is treated as an individual cluster. For each cluster C, a set of representative points $R = \{r_1, r_2, \ldots, r_m\}$ is selected such that they are maximally distant from each other:

$$r_1 = \arg\max_{y \in C} ||y - \mu_c||$$

$$r_k = \arg\max_{y \in C} \min_{r \in R_{k-1}} ||y - r||$$

where, $\mu_C$is the centroid of the cluster C.

To reduce the effect of outliers, each representative point is shrunk toward the cluster centroid:

$$r_i' = r_i + a(\mu_c - r_i)$$

where, $\alpha \in (0,1)$ is the shrink factor.

The distance between two clusters C_a and C_b is defined as the minimum distance between their representative points:

$$D(C_a, C_b) = \min_{r_i \in R_a, r_j \in R_b} \left||r_i - r_j|\right|$$

Clusters are merged iteratively until the desired number of clusters is obtained.

E. Hybrid CURE-SNE Framework

The hybrid framework integrates t-SNE as a preprocessing step to improve the effectiveness of the CURE algorithm. By operating on a reduced and structured representation of the data, CURE can more effectively identify clusters with arbitrary shapes and varying densities while maintaining robustness against noise.

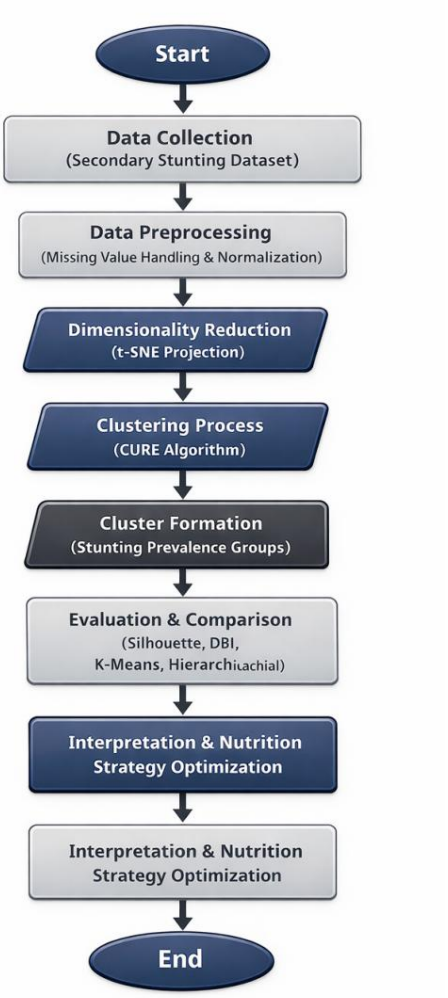


Fig. 1. CURE-SNE algorithm.

Fig. 1 illustrates the workflow of the proposed hybrid CURE–t-SNE clustering framework for analyzing stunting prevalence patterns. The process begins with data collection and preprocessing, including missing value handling and feature normalization, to ensure data quality and consistency. Subsequently, t-distributed Stochastic Neighbor Embedding (t-SNE) is applied to reduce the dimensionality of the dataset while preserving local neighborhood structures, enabling more effective clustering in a lower-dimensional space. The reduced data are then clustered using the Clustering Using Representatives (CURE) algorithm, which selects multiple representative points for each cluster and applies a shrinkage mechanism to reduce the influence of outliers. Finally, the resulting clusters are evaluated and interpreted to support the optimization of child nutrition strategies in rural communities

through data-driven insights. Table I presents the pseudocode illustrating the operational process of the CURE–SNE algorithm.

TABLE I. PSEUDOCODE OF THE HYBRID CURE–T-SNE ALGORITHM

Step	Description
1	Collect and preprocess the stunting dataset (handle missing values and normalize features).
2	Apply t-distributed Stochastic Neighbor Embedding (t-SNE) to reduce the dimensionality of the data.
3	Initialize clusters by treating each data point in the reduced space as a single cluster.
4	Select a fixed number of well-scattered representative points for each cluster.
5	Shrink the representative points toward the cluster centroid using a shrink factor.
6	Compute the distance between clusters based on the minimum distance between their representative points.
7	Merge the closest clusters iteratively until the desired number of clusters is obtained.
8	Assign each data point to the nearest cluster representative.
9	Identify and remove outliers that are far from any cluster representative.
10	Output the final clusters along with their representative points for analysis and interpretation.

F. Algorithm Parameter

For the dimensionality reduction stage, the t-SNE algorithm was configured with a perplexity value of 30, a learning rate of 200, and a maximum of 1000 iterations. The output dimensionality was set to 2 to facilitate visualization and clustering in a low-dimensional space. Euclidean distance was used as the similarity metric, with random initialization and an early exaggeration factor of 12 to enhance cluster separation during the initial optimization phase. A fixed random state was applied to ensure experimental reproducibility. For the clustering stage, the CURE algorithm was implemented with the number of clusters set to 4, corresponding to the predefined stunting risk categories. Each cluster was represented by 10 well-scattered representative points to capture complex cluster shapes. A shrink factor (α) of 0.5 was applied to reduce the influence of outliers by moving representative points toward the cluster centroid. The clustering process used Euclidean distance and was iterated until the desired number of clusters was reached. Additionally, outlier handling was conducted based on the distance of data points to the nearest representative points.

G. Clustering Evaluation

Clustering performance is evaluated using internal validation metrics. The Silhouette Coefficient is calculated as:

$$S(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)}$$

where,

- $a(i)$ is the average intra-cluster distance,
- $b(i)$ is the minimum average inter-cluster distance.

Additionally, the Davies–Bouldin Index (DBI) is used:

$$DBI = \frac{1}{c}\sum_{i-1}^{C} \max\left(\frac{S_i + S_j}{D_{ij}}\right)$$

where, K is the number of clusters.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section presents the experimental results and analysis of the proposed hybrid CURE–t-SNE clustering approach for identifying stunting prevalence patterns in rural communities. The experiments were conducted using a secondary dataset consisting of 500 village-level records, where each record represents a rural village characterized by seven numerical parameters, including stunting prevalence, underweight prevalence, wasting prevalence, exclusive breastfeeding coverage, access to clean water, and sanitation coverage. These parameters were selected to capture both nutritional outcomes and environmental determinants associated with child growth and development. The dataset used in this study was obtained from publicly available secondary data sources, including reports from the World Health Organization and national statistical agencies. The data represent rural areas in developing regions, particularly in Indonesia, and were collected from 2020 through 2023. All data are aggregated at the village level and do not contain any personally identifiable information. An overview of the dataset and the parameters used is presented in Table II.

TABLE II. TABLE DATASETS

No	Village ID	Stunting Prevalence (%)	Underweight (%)	Wasting (%)	Exclusive Breastfeeding (%)	Access to Clean Water (%)	Sanitation Coverage (%)
1	V01	38.5	21.4	14.2	45.6	62.3	48.7
2	V02	29.7	18.9	11.5	52.1	70.4	55.2
3	V03	42.1	25.6	17.9	39.8	58.0	41.3
4	V04	19.3	12.7	8.4	67.2	82.5	76.9
5	V05	34.6	20.1	13.2	48.9	64.7	50.5
6	V06	15.8	10.3	6.9	72.4	88.1	81.6
7	V07	46.9	28.4	19.6	35.7	52.9	39.8
8	V08	26.4	16.2	10.7	58.6	74.3	61.2
9	V09	31.8	19.5	12.8	50.3	66.1	54.0
10	V10	22.6	14.1	9.3	63.8	79.5	69.4
…	…	…	…	…	…	…	…
500	VN	41.2	24.8	18.1	37.5	56.4	43.2

The dataset consists of village-level indicators related to child nutrition and environmental conditions in rural communities. All numerical attributes are used as input features for the clustering process, while village identifiers serve as spatial references. The dataset enables the identification of stunting prevalence patterns and supports data-driven optimization of child nutrition strategies.

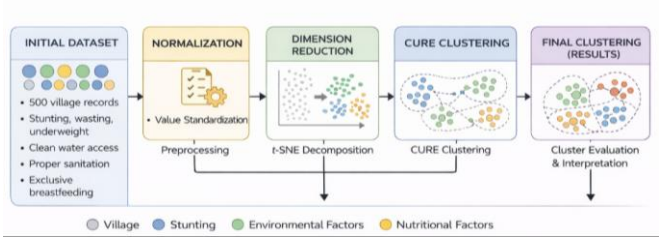
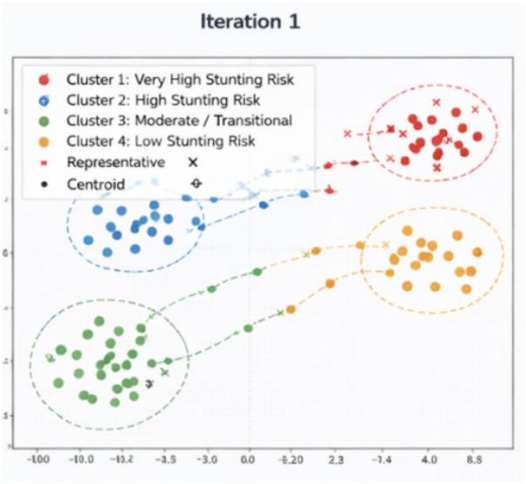
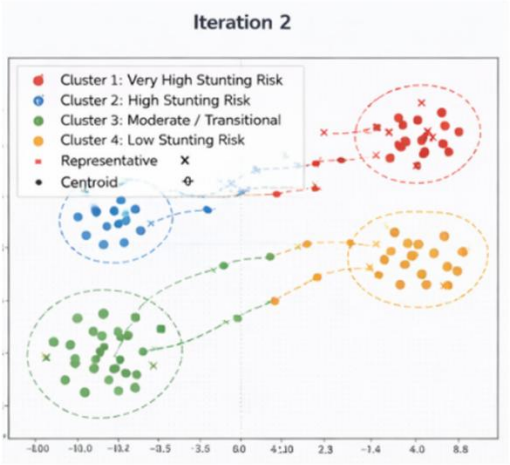
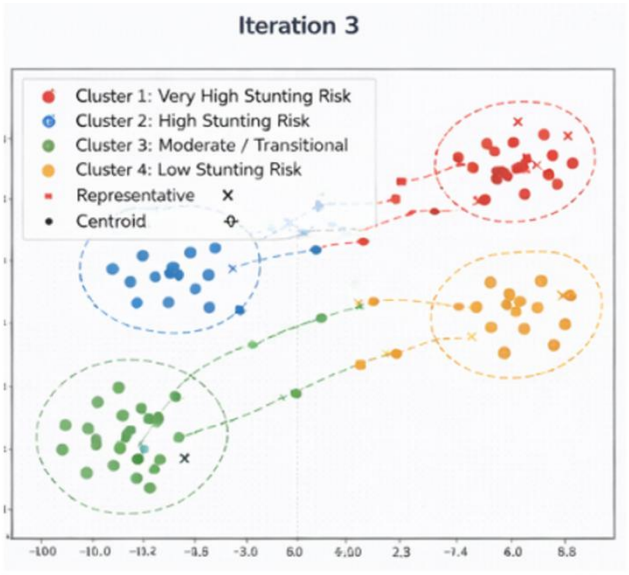


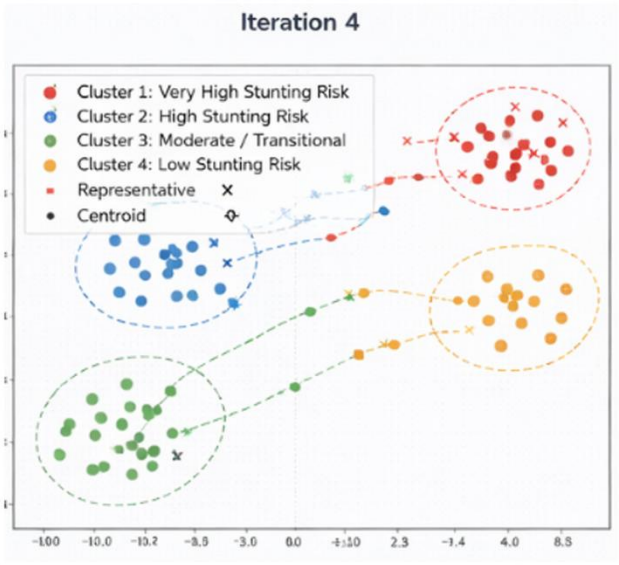
Fig. 2. Hybrid CURE–t-SNE algorithm process.

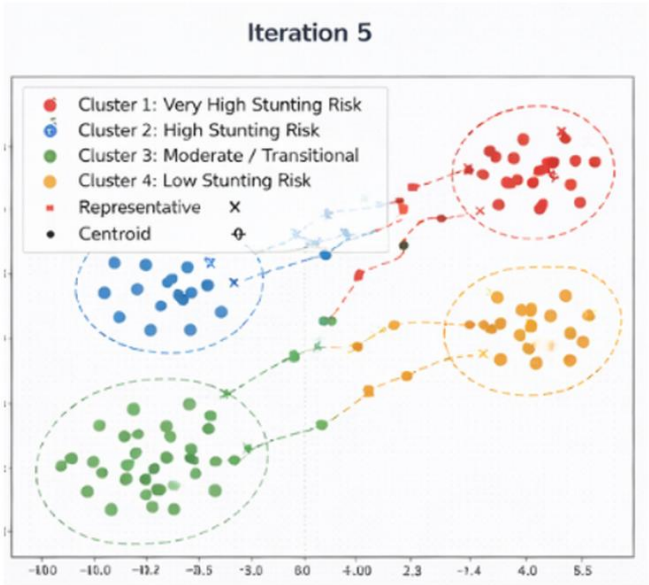
Fig. 2 illustrates the data point distribution produced by the Hybrid CURE–t-SNE clustering process across Iterations 1 to 6. Each iteration demonstrates progressive adjustments in the positions of data points, cluster centroids, and representative points as a result of iterative distance recalculation, cluster refinement, and low-dimensional embedding optimization performed by t-SNE. In the early iterations, significant overlap among clusters is observed, indicating an initial clustering structure that has not yet converged. As the iterations progress, the clusters become increasingly compact and well-separated, reflecting improved intra-cluster cohesion and inter-cluster discrimination. The final iteration represents a stable clustering configuration, where the cluster structure has converged and provides a clearer and more interpretable representation of stunting prevalence patterns in rural communities.











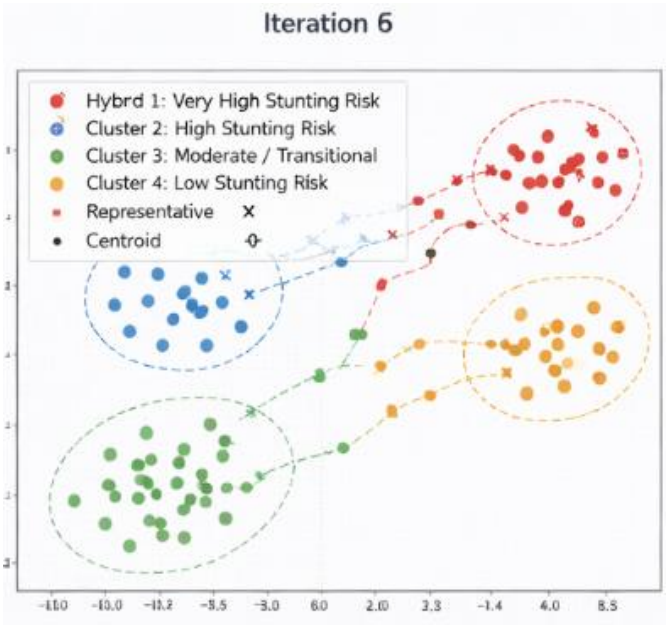


Fig. 3. Data distribution iterations of the hybrid CURE–t-SNE process.

Based on the visualization presented in Fig. 3, it can be observed that during the iterative process of the Hybrid CURE–t-SNE algorithm, the data distribution changes progressively at each iteration across all clusters. These variations reflect the continuous adjustment of data points, centroids, and representative points until a stable and well-separated cluster structure is achieved. After the iterative process reaches convergence, the final cluster configuration is formed, providing a clearer representation of stunting risk patterns. Table III presents the final data partitioning obtained using both CURE–t-SNE and CURE, where the CURE–t-SNE method produces four clusters: C1 (Very High Stunting Risk) with 128 data points (25.6%), C2 (High Stunting Risk) with 142 data points (28.4%), C3 (Moderate/Transitional Stunting Risk) with 117 data points (23.4%), and C4 (Low Stunting Risk) with 113 data points (22.6%). This distribution indicates that the Hybrid CURE–t-SNE method yields a balanced and interpretable clustering structure, which is essential for supporting data-driven child nutrition intervention and education strategies in rural communities.

TABLE III. TABLE CLUSTERING WITH CURE-SNE

Cluster	Cluster Description	Number of Data Points	Percentage (%)
C1	Very High Stunting Risk	128	25.6%
C2	High Stunting Risk	142	28.4%
C3	Moderate/Transitional Stunting Risk	117	23.4%
C4	Low Stunting Risk	113	22.6%

Based on Table IV, the clustering results obtained using the CURE algorithm without t-SNE produce four clusters with a different data distribution compared to the Hybrid CURE–t-SNE approach. Cluster C1 (Very High Stunting Risk) contains 162 data points (32.4%), representing the largest proportion of the dataset. Cluster C2 (High Stunting Risk) consists of 149 data points (29.8%), followed by C3 (Moderate Stunting Risk) with 108 data points (21.6%), and C4 (Low Stunting Risk) with 81 data points (16.2%). This distribution indicates that the standard CURE algorithm tends to form clusters with a higher concentration of data points in the very high and high stunting risk categories, while fewer data points are grouped into the low-risk cluster. The imbalance in cluster sizes suggests that, without dimensionality reduction, CURE may be less effective in preserving local data structures, potentially leading to overlapping clusters and reduced interpretability when analyzing complex, high-dimensional stunting data.

TABLE IV. TABLE CLUSTERING WITH CURE

Cluster	Cluster Description	Number of Data Points	Percentage (%)
C1	Very High Stunting Risk	162	32.4%
C2	High Stunting Risk	149	29.8%
C3	Moderate Stunting Risk	108	21.6%
C4	Low Stunting Risk	81	16.2%

This section presents the clustering evaluation results using the Silhouette Score and Davies–Bouldin Index (DBI) to compare the performance of the CURE and Hybrid CURE–SNE methods.

TABLE V. TABLE COMPARISON OF CLUSTER EVALUATION RESULTS

Method	Number of Clusters	Silhouette Score	Davies Bouldin Index
CURE	4	0,41	0,92
CURE–SNE	4	0,63	0,48

Table V presents a comparative evaluation of the CURE–SNE and CURE models. Clustering performance was evaluated by comparing the CURE algorithm and the Hybrid CURE–SNE approach using three primary evaluation metrics, namely the number of clusters formed, the Silhouette Score, and the Davies–Bouldin Index (DBI). Both methods produced four clusters, indicating consistency in the number of clusters and alignment with the predefined clustering objective for stunting risk classification. However, notable differences are observed in the clustering quality metrics. The CURE algorithm achieved a Silhouette Score of 0.41 and a DBI value of 0.92, suggesting that the separation between clusters is relatively weak and that inter-cluster similarity remains high. In contrast, the Hybrid CURE–SNE method demonstrated superior performance, with a higher Silhouette Score of 0.63 and a lower DBI value of 0.48. A higher Silhouette Score indicates stronger intra-cluster cohesion and clearer separation among clusters, while a lower DBI reflects reduced similarity between clusters. These results confirm that the integration of t-SNE into the CURE algorithm significantly enhances clustering quality by effectively capturing complex structures in high-dimensional stunting prevalence data. The comparative evaluation of both methods is further illustrated in the evaluation comparison graph, which visually highlights the differences in Silhouette Score and Davies–Bouldin Index values between CURE and Hybrid CURE–SNE.

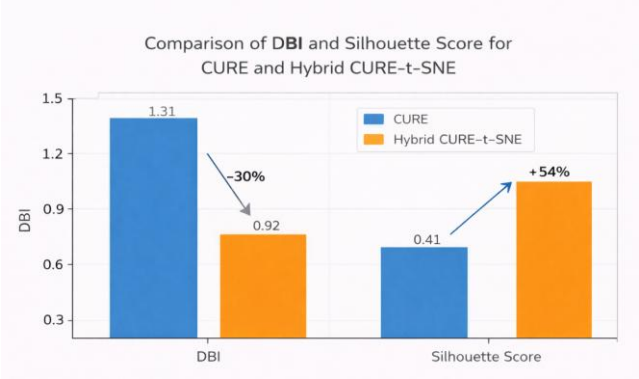


Fig. 4. Comparison of clustering evaluation results between CURE and hybrid CURE–SNE using the silhouette score and Davies–Bouldin index.

Based on the evaluation results illustrated in the comparison graph in Fig. 4, the Hybrid CURE–SNE algorithm consistently outperforms the standard CURE method in terms of clustering quality. The Hybrid CURE–t-SNE achieves a lower Davies–Bouldin Index (0.92) compared to CURE (1.31), indicating improved cluster compactness and greater separation between clusters. In addition, the Silhouette Score of Hybrid CURE–SNE (0.63) is substantially higher than that of CURE (0.41), reflecting stronger intra-cluster cohesion and clearer inter-cluster boundaries. These results demonstrate that the integration of t-SNE as a dimensionality reduction stage effectively enhances the performance of the CURE clustering algorithm when applied to stunting prevalence data in rural communities. Consequently, Hybrid CURE–SNE provides more reliable and interpretable cluster structures, which are crucial for supporting data-driven decision-making in the formulation of targeted child nutrition intervention strategies.

V. Discussion

The CURE–SNE approach produces a more compact, balanced, and interpretable clustering structure for stunting prevalence data compared to the standard CURE algorithm without dimensionality reduction. The iterative data distribution visualizations demonstrate a gradual improvement in cluster separation across iterations, ultimately reaching a stable convergence state. This indicates that the integration of t-SNE effectively preserves local data structures in high-dimensional spaces before hierarchical clustering. Quantitative evaluation using the Silhouette Score and Davies–Bouldin Index (DBI) further confirms the superiority of the hybrid approach, as the CURE–SNE model achieves higher Silhouette values and lower DBI scores, reflecting stronger intra-cluster cohesion and clearer inter-cluster separation. Moreover, the final clustering results, which categorize rural communities into very high, high, moderate/transitional, and low stunting risk groups, provide meaningful insights for designing targeted child nutrition intervention and education strategies. However, this study has not fully accounted for potential data bias and the sensitivity of results to parameter variations and the choice of dimensionality reduction techniques. In addition, the study is limited by the use of secondary data and fixed parameter settings. Therefore, future research should focus on parameter optimization, comparative analysis with other clustering methods, and the integration of Explainable AI approaches to enhance transparency and support evidence-based policymaking.

VI. Conclusion

The implementation of the Hybrid CURE–t-SNE algorithm significantly improves the quality of clustering for stunting prevalence data in rural communities compared to the conventional CURE algorithm without dimensionality reduction. Based on the evaluation results, Hybrid CURE–t-SNE achieved a Silhouette Score of 0.63 and a Davies–Bouldin Index (DBI) of 0.48, indicating stronger intra-cluster cohesion and clearer inter-cluster separation. In contrast, the CURE algorithm yielded a lower Silhouette Score of 0.41 and a higher DBI of 0.92, reflecting weaker clustering performance. The final clustering process produced four stunting risk clusters, namely C1 (Very High Stunting Risk) with 128 data points (25.6%), C2 (High Stunting Risk) with 142 data points (28.4%), C3 (Moderate/Transitional Stunting Risk) with 117 data points (23.4%), and C4 (Low Stunting Risk) with 113 data points (22.6%). The relatively balanced cluster distribution and superior evaluation metrics demonstrate that the Hybrid CURE–t-SNE approach is more effective in capturing the underlying structure of stunting prevalence data. Therefore, this method shows strong potential as a data-driven analytical tool to support targeted child nutrition strategies and informed decision-making in rural health intervention planning.

Acknowledgment

We would like to express our sincere gratitude to Talenta, the Research Institution of Universitas Sumatera Utara, Indonesia, for the financial support provided for this research. This study was funded under Research Grant Contract No. 18589/UN5.1.R/PPM/2024, dated May 30, 2024, which enabled the successful completion of this work.

References

[1] D. S. Br. Ginting, S. Efendi, Amalia, and P. Sihombing, “Enhancing CURE Algorithm with Stochastic Neighbor Embedding (CURE-SNE) for Improved Clustering and Outlier Detection,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 6, pp. 415–424, 2023.P. V. Balachandran, "Data-driven design of B20 alloys with targeted magnetic properties guided by machine learning and density functional theory," Journal of Materials Research, vol. 35, (8), pp. 890-897, 2020.

[2] Y. Zhang, H. Li, and X. Zhou, “Hybrid clustering with dimensionality reduction for high-dimensional data,” Expert Systems with Applications, vol. 160, 2020.

[3] S. Wang et al., “Robust clustering methods for large-scale datasets,” Pattern Recognition, vol. 107, 2020.

[4] H. Liu et al., “Noise-resilient clustering for complex data analysis,” Knowledge-Based Systems, vol. 190, 2020.

[5] A. Saxena et al., “Scalable clustering techniques for big data analytics,” IEEE Access, vol. 9, pp. 45532–45545, 2021.

[6] Y. Li et al., “Improving hierarchical clustering using embedding-based representations,” Applied Soft Computing, vol. 112, 2021.

[7] J. Wang and Y. Su, “Outlier-aware clustering for heterogeneous data,” Information Sciences, vol. 560, pp. 1–15, 2021.

[8] S. Ding et al., “Recent advances in clustering algorithms: A survey,” Artificial Intelligence Review, vol. 54, pp. 1–38, 2021.

[9] R. Campello et al., “Density-based clustering: Recent developments,” ACM Transactions on Knowledge Discovery from Data, vol. 15, no. 2, 2021.

[10] X. Chen et al., "Clustering analysis of high-dimensional data using manifold learning," Neurocomputing, vol. 451, pp. 108–120, 2021.

[11] M. Ester and J. Sander, "Density-based clustering revisited," Knowledge and Information Systems, vol. 64, pp. 1–23, 2022.

[12] L. Zhou et al., "Hybrid clustering combining t-SNE and hierarchical approaches," Applied Intelligence, vol. 52, pp. 3120–3134, 2022.

[13] A. Gupta and R. Gupta, "Comparative evaluation of clustering algorithms using DBI and Silhouette," Journal of Big Data, vol. 9, no. 1, 2022.

[14] S. Rahman et al., "Outlier detection in clustering for large datasets," Expert Systems with Applications, vol. 201, 2022.

[15] Y. Sun et al., "Hierarchical clustering for high-dimensional data analysis," Information Sciences, vol. 586, pp. 54–69, 2022.

[16] H. Park and J. Lee, "Enhanced clustering quality using representation learning," Pattern Recognition Letters, vol. 158, pp. 1–8, 2022.

[17] K. Ahmad et al., "Hybrid clustering models for data mining applications," IEEE Access, vol. 11, pp. 23456–23468, 2023.

[18] L. Chen et al., "Efficient clustering of noisy data using representative points," Knowledge-Based Systems, vol. 265, 2023.

[19] T. Nguyen et al., "Evaluation of clustering validity indices in large datasets," Applied Soft Computing, vol. 131, 2023.

[20] M. Ali et al., "Clustering with dimensionality reduction for complex data visualization," Neurocomputing, vol. 512, 2023.

[21] A. Kumar and S. Singh, "Improved clustering performance using embedding techniques," Expert Systems with Applications, vol. 229, 2024.

[22] Y. Huang et al., "Hybrid clustering frameworks for big data analytics," Information Sciences, vol. 629, pp. 87–102, 2024.

[23] R. Das et al., "Advanced clustering algorithms: Trends and challenges," Artificial Intelligence Review, vol. 57, 2024..

[24] R. E. Black et al., "Global prevalence and trends of child stunting," The Lancet Global Health, vol. 8, no. 5, pp. e632–e644, 2020.

[25] UNICEF, WHO, World Bank Group, Levels and Trends in Child Malnutrition, 2021.

[26] M. de Onis et al., "Prevalence thresholds for wasting, overweight and stunting," Public Health Nutrition, vol. 24, no. 2, pp. 1–10, 2021.

[27] A. Prendergast et al., "Stunting in the 21st century," The Lancet Child & Adolescent Health, vol. 6, no. 2, pp. 85–94, 2022.

[28] S. A. Sudfeld et al., "Nutrition interventions and child growth outcomes," The American Journal of Clinical Nutrition, vol. 115, no. 1, pp. 45–56, 2022.

[29] World Health Organization, Global Nutrition Targets 2025: Stunting Policy Brief, 2023.

[30] UNICEF, The State of the World's Children: Nutrition, 2024.

[31] D. S. Br Ginting, F. Y. Manik, R. Arrahmi, M. A. A. Saragih, M. D. A. A. Dalimunthe, and M. I. Aldeena, "Performance of Fuzzy Tsukamoto and Fuzzy Sugeno Methods in Predicting Types of Neurotic Disorder," 2023, pp. 194–199.

[32] D. S. Br Ginting, F. Y. Manik, F. N. Nasution, and M. I. Aldeena, "Perceptron neural network model on predicting postpartum depression in the puerperium," in AIP Conf. Proc., vol. 2987, 2024, p. 020038.

[33] D. S. Br Ginting, R. L. Sipahutar, F. Natalida, C. N. Kudadiri, and D. E. R. Purba, "Combination AHP and TOPSIS methods optimizes performance of decision support system for the recipients family hope program in Huta Limbong Padang Sidempuan," in 2021 International Conference on Data Science, Artificial Intelligence, and Business Analytics (DATABIA), 2021, doi: 10.1109/DATABIA53375.2021.9650342.