Hybrid Genetic Algorithm and Particle Swarm Optimization as Centroid Initialization in K-Means Clustering for National Nutrition Dataset

Maman Somantri [1,*], Reza Iqbal Pramudya[2], Aris Sugiharto[3] & Tonni Agustiono Kurniawan[4]

[1]Department of Computer Engineering, Diponegoro University,
Jalan Prof. H. Soedarto, S.H., Semarang 50275, Central Java, Indonesia
[2]School of Postgraduate Studies, Diponegoro University,
Jalan Imam Bardjo SH No. 3-5, Semarang 50241, Central Java, Indonesia
[3]Department of Informatics, Diponegoro University,
Jalan Prof. H. Soedarto, S.H., Semarang 50275, Central Java, Indonesia
[4]College of Ecology and Environment, No. 4221 South Xiang'an Road, Xiang'an District, Xiamen City, Fujian Province, 361102, P.R. China
*E-mail: mmsomantri@live.undip.ac.id

Abstract. The efficacy of large-scale public health interventions, such as Indonesia's Makan Bergizi Gratis (MBG) program, relies heavily on the precise identification of vulnerable regions. Conventional clustering techniques like K-means are essential for such mapping but frequently suffer from performance degradation due to sensitivity to random centroid initialization, leading to suboptimal and unstable solutions. This study proposes a sequential hybrid metaheuristic framework (K-means+GAPSO) that synergizes the broad global exploration of genetic algorithms (GA) with the rapid local exploitation of particle swarm optimization (PSO). The proposed model was rigorously validated using the Indonesian National Nutrition Dataset (2021–2023) through 10 independent runs to ensure stochastic robustness. Computational results at the optimal cluster count (K = 3) demonstrated that the K-means+GAPSO (prob = 0.3) configuration significantly outperformed the standard baselines, achieving a highly stable silhouette mean of 0.7130 (±0.0036) and a Davies-Bouldin index mean of 0.4708 (±0.0178). This metric achievement represents a substantial performance improvement in separation and compactness compared to standard K-means. The implementation of this structurally robust method for nutritional clustering provides a reliable analytical foundation for policymakers to accurately target food security initiatives and minimize resource misallocation.

Keywords: *genetic algorithm; hybrid clustering; K-means; national nutrition; particle swarm optimization.*

1 Introduction

The strategic implementation of the Free Nutritious Meal (*Makan Bergizi Gratis - MBG*) program represents a cornerstone of Indonesia's national commitment to

Received on December 8th, 2025, 1st Revision on April 17th, 2026, 2nd Revision on May25th, 2026, Accepted for publication June 23rd, 2026.

eradicating malnutrition and preparing for the 'Golden Generation 2045.' As a massive public health intervention, the success of MBG relies heavily on precise targeting to ensure that resources effectively reach the most vulnerable populations across Indonesia's vast and heterogeneous archipelago. Inaccurate mapping of nutritional pockets risks misdirecting vital state resources, potentially undermining the program's core objective of health equity. Consequently, the ability to process complex national health data to guide these interventions is a prerequisite for effective governance [1]. In this context, the role of advanced computational analysis becomes paramount [2].

The rapid expansion of digital information has made data mining a critical necessity for extracting strategic insight into decision-making processes [3]. Among the foundational techniques in unsupervised learning, clustering occupies a central position, offering the capability to organize unlabeled data into meaningful groups based on intrinsic similarities [4]. This process is immensely valuable across diverse fields, ranging from logistics and image processing to sophisticated medical analysis [5].

The K-means algorithm remains one of the most widely adopted partitioning methods due to its procedural simplicity and computational efficiency [6]. K-means fundamentally operates by partitioning n data points into k clusters, striving to minimize the sum of squared errors (SSE) or intra-cluster variance. However, this popularity belies a significant inherent flaw: K-means is highly sensitive to the initial selection of cluster centroids [3]. A random or suboptimal starting point can cause the algorithm to swiftly converge to a local optimum (*optima lokal*) instead of finding the best global cluster configuration. This instability results in unreliable and suboptimal clustering outcomes, particularly when dealing with complex or high-dimensional datasets. Although improvements such as K-means++ offer a 'smarter' initialization method, they still rely on a heuristic approach and do not guarantee a globally optimal solution [4].

To overcome this persistent challenge, researchers have frequently turned to metaheuristic optimization algorithms that possess robust global search capabilities [3]. This study leverages the synergistic combination of two prominent population-based approaches: the genetic algorithm (GA) and particle swarm optimization (PSO). The genetic algorithm, inspired by Darwinian evolution [7], excels at global exploration [4], using selection, crossover, and mutation operators to broadly sweep the solution space [8] and maintain population diversity [7], thereby effectively avoiding local optima traps [9]. Conversely, particle swarm optimization, modeled after swarm intelligence, is renowned for its rapid convergence speed and superior capability for local exploitation, guiding particles towards the best global position (*gbest*) found so

far. By integrating these two techniques, Hybrid GAPSO achieves a crucial balance between exploration and exploitation, leading to demonstrably superior optimization results compared to using a single algorithm [4].

The practical application of this enhanced methodology targets a critical domain: national health policy. This research proposes implementing the Hybrid GAPSO model to optimize centroid initialization for K-means clustering [10] on the Indonesian National Nutrition Dataset (Data Gizi Indonesia: Prevalensi dan Konsumsi) for 2021–2023 [18]. This multidimensional dataset, comprising 1,542 data entries, includes vital public indicators such as protein and energy consumption, and the prevalence of undernourishment. Misclassification or inaccurate grouping of regional nutritional profiles carries serious consequences, including inefficient resource allocation and the implementation of ill-targeted interventions [3].

Exploration of clustering optimization based on hybrid GA-PSO architectures has previously been extensively investigated theoretically. The methodological novelty of this study specifically lies in the integration of the algorithmic architecture with a pure Winsorization preprocessing method and a variance inflation factor (VIF) based feature selection mechanism, strictly without involving dimensionality scaling that distorts natural spatial distances. This comprehensive experimental approach empirically proves capable of producing absolute convergence of evaluation metrics on demographic-spatial datasets. The achievement of such computational stability, characterized by zero metric variance in a stochastic modeling environment, represents a highly rare robustness trait in conventional hybrid algorithm implementations.

Therefore, this paper aims to implement the Hybrid Genetic Algorithm and Particle Swarm Optimization (GAPSO) model for centroid initialization in K-means clustering. This approach is expected to generate more accurate, stable, and representative cluster results for Indonesia's nutritional status compared to conventional methods. The effectiveness and quality of the resulting clusters will be rigorously assessed using key internal validation metrics: the silhouette score and the Davies-Bouldin index (DBI).

2 Relevant Work

The pervasive adoption of K-means clustering across fields ranging from logistics to medical diagnostics is driven by its computational simplicity and efficiency. However, the core vulnerability of K-means, its extreme sensitivity to the random initialization of cluster centroids, remains a significant challenge. A suboptimal starting configuration frequently results in the algorithm converging prematurely to a local optimum rather than identifying the optimal global partitioning of the

data [11]. Although variations such as K-means++ provide a heuristically 'smarter' starting point, this approach still offers no guarantee of achieving the global optimum, leaving the door open for novel initialization strategies [10].

To address this persistent instability, recent research has increasingly turned to metaheuristic optimization algorithms that possess robust global search capabilities [3]. This study focuses on the synergistic strength of combining a genetic algorithm and particle swarm optimization. GA, inspired by natural selection, is highly effective for global exploration and maintaining population diversity, utilizing operators like crossover and mutation to avoid local optimum traps. Conversely, PSO, derived from swarm intelligence, is lauded for its fast convergence and superior capacity for local exploitation, quickly refining promising solutions. Crucially, while GA tends toward slow convergence and PSO is prone to premature local optima when used in isolation, their hybrid combination (GAPSO) leverages GA's exploration to feed PSO's exploitation power, achieving the superior balance necessary for complex optimization tasks [7].

While hybrid metaheuristic algorithms, including GA-PSO frameworks, have been extensively applied to optimize clustering, a critical review of the literature reveals a significant application gap. Previous studies have primarily deployed these frameworks in domains such as distributed task scheduling [7], document clustering [10], or logistics route optimization [12]. In these contexts, K-means is frequently reduced to a secondary sub-process rather than the primary subject of optimization. Furthermore, studies utilizing algorithms like SA-PSO-GK++ or GA-based feature selection often rely on datasets with relatively well-behaved, continuous distributions, such as standard image processing or benchmark medical datasets [13].

These existing approaches rarely address the unique topological challenges inherent in public-health and socioeconomic datasets. In the context of national food security and public health analytics, clustering is a critical instrument for policy formulation and targeted intervention, frequently used to categorize regions based on macronutrient consumption and the prevalence of undernourishment (PoU) indicator [11], [14]. However, such national data is intrinsically complex featuring extreme regional disparities, massive scale differences, and non-Gaussian distributions with natural outliers. Straightforward applications of standard K-means or isolated hybrid algorithms often suffer from premature convergence or instability when forced to navigate the sparse, skewed spaces typical of real-world nutritional profiling. Misidentifying a region's nutritional status due to suboptimal clustering can result in severely misallocated government resources.

Therefore, implementing a robust optimization framework is not merely a computational enhancement but a practical public health necessity. The methodological novelty of this study lies in the implementation of a pure, sequential GA-to-PSO injection framework specifically engineered to solve the fundamental centroid initialization flaw of K-means in highly volatile, multi-dimensional nutritional datasets [3]. Unlike standard K-means++ or isolated PSO methods that struggle with such volatile data topologies, this sequential hybridization leverages GA to broadly explore the skewed spatial boundaries without eliminating critical outliers. This global exploration is immediately followed by PSO for precise local exploitation [4]. Ultimately, this study bridges the gap between algorithmic optimization and public health analytics, providing a rigorously validated methodology using the silhouette score and the Davies-Bouldin index (DBI) to derive superior, stable clusters for evidence-based policy intervention [15].

3 Methodology

The standard K-means clustering algorithm suffers from a critical structural flaw: its extreme sensitivity to the initial placement of cluster centroids. This instability often forces the algorithm to prematurely converge at a local optimum, which is unacceptable when deriving insights from crucial public policy data, such as national nutrition profiles. To address this persistent limitation and ensure a superior global clustering solution, we propose the Hybrid Genetic Algorithm and Particle Swarm Optimization for K-means (GAPSO-K-means) model, specifically designed to calculate the optimal initial centroid positions.

3.1 Data Description

The proposed methodology was validated using the Indonesian Nutrition Dataset (2021–2023), which comprises 1,542 non-null instances. As detailed in Table 1, the dataset's 12 columns (10 numerical, 2 categorical) were complete, obviating the need for data imputation. An Exploratory Data Analysis (EDA) was first conducted to examine distributions and correlations, informing feature selection. Subsequently, model evaluation using silhouette score and Davies-Bouldin index (DBI) to measure cluster cohesion. Separation confirmed the hybrid model's capability to deliver stable and accurate clustering outcomes [11].

Table 1 Data summary.

Column	Non-Null Count	Dytpe
Regency City	1542 non-null	object
Province	1542 non-null	object
Year	1542 non-null	int64
Protein Consumption	1542 non-null	float64
PoU	1542 non-null	float64
Total Population	1542 non-null	int64
Undernourished Inhabitants	1542 non-null	int64
Food Security Index	1542 non-null	float64
District/City Ranking	1542 non-null	int64
Food Security Index Group	1542 non-null	int64
Hope Food Pattern Score	1542 non-null	float64
Energy Consumption	1542 non-null	int64

Table 2 presents a comprehensive descriptive statistical summary of the numerical features, which was performed to ascertain their underlying characteristics and distributions. The analysis covers data spanning the period from 2021 to 2023.

Table 2 Descriptive statistics.

	count	mean	std	min	25%	50%	75%	max
Year	*1542.0*	2022.000000	0.816761	2021	2021.0000	2022.0000	2023.0000	2023.00
Protein Consumption PoU	*1542.0*	60.510117	8.760166	21.60	56.3000	60.600	65.0000	109.10
Total Population	*1542.0*	12.587231	10.059409	0.65	6.7300	9.835	14.3775	80.32
Undernourish Inhabitants	*1542.0*	536378.243191	637199.152287	22926.00	158453.5000	292297.000	658888.5000	5627021.00
Food Security Index	*1542.0*	50059.441634	51040.047511	675.00	15793.5000	30518.000	65783.7500	390608.00
District/City Ranking	*1542.0*	72.802633	14.592854	14.54	68.6725	77.065	82.3350	95.80
Food Security Index Group	*1542.0*	178.184825	125.441151	1.00	65.0000	159.500	288.0000	416.00
Hope Food Pattern Score	*1542.0*	5.086900	1.406473	1.00	5.0000	6.000	6.0000	6.00
Energy Consumption	*1542.0*	83.199157	8.825659	39.60	78.0000	84.700	90.0000	99.50

The analysis in Table 2 shows significant variability and, most critically, profound disparities in scale across features; for instance, 'Total population' is measured in millions while 'PoU' is measured in tens. This vast difference in magnitudes provides a robust justification for mandatory feature scaling (e.g., Standardization and Min-Max) as a preprocessing step. This is crucial because distance-based algorithms like K-means and GA are highly sensitive to data scale and scaling ensures all variables contribute equitably to the cluster formations.

3.2 Data Preprocessing and Feature Selection

The data preprocessing pipeline in this study strictly employed a pure Winsorization technique combined with a variance inflation factor (VIF) based feature selection. The complete omission of any spatial-distorting dimensionality

scaling methods preserves the original coordinate integrity of the nutritional variables. A rigorous preprocessing pipeline was applied following the exploratory data analysis findings to ensure algorithmic stability and prevent bias in the distance-based clustering computations [16]. The categorical attribute 'Province' was deliberately excluded from the computational input during feature selection. The incorporation of categorical nominal data into K-means requires techniques like One-Hot Encoding, which would artificially bias the algorithm to cluster regions based on geographical boundaries [17]. The 'City/Regency' attribute was also excluded from the computations and retained exclusively as a post-clustering identifier. A methodological note regarding the spatial feature 'Space' dictates that the 'Year' variable was strictly excluded from the computational training process. The temporal attribute was not involved in any mathematical Euclidean distance calculations during the clustering execution. The retention of this specific variable within the dataset structure serves exclusively as an identity index to track longitudinal observations across the evaluated period, ensuring that the integrity of the spatial computations is solely determined by genuine nutritional indicators.

Significant variability and extreme outliers were observed across several continuous variables, particularly in Population and the prevalence of undernourishment (PoU) indicator. The presence of such extreme natural disparities can severely distort Euclidean distance calculations and trap standard clustering algorithms in local optima. A robust Winsorization technique was strictly applied to systematically mitigate this disruptive mathematical impact without entirely eliminating the critical signals of highly vulnerable regions. This transformation effectively capped the extreme outlier thresholds, substituting values beyond the 5th and 95th percentiles with the respective boundary values. A variance inflation factor evaluation was systematically conducted to eliminate multicollinearity among the nutritional indicators and further refine the feature space. The application of Winsorization fundamentally preserved the underlying distributional integrity of the national nutrition data while establishing a stabilized spatial topology. This stabilized environment acts as a crucial computational prerequisite for the GA and PSO framework to conduct an unbiased and globally optimal spatial exploration without the necessity of aggressive data scaling.

3.3 Hybrid GAPSO Optimization Framework

The central element of this research is the two-stage GAPSO model, which optimizes centroid initialization to maximize the cluster quality.

1. Phase 1: Global exploration via genetic algorithm: The GA is executed first to perform a wide-ranging global exploration of the solution space. By utilizing selection, crossover, and mutation operators, the GA effectively

avoids the pitfalls of local optima and identifies regions containing the most promising candidate centroids.
2. Phase 2: Local exploitation using particle swarm optimization: The best set of centroids (the fittest chromosome) identified by the GA is used to initialize the PSO swarm. PSO then performs rapid local exploitation (refinement), iteratively adjusting particle positions based on pBest and gBest to quickly converge on the optimal configuration. This step ensures a superior balance between exploration and exploitation.

Refinement rate tuning: The PSO process is tuned by varying its refinement rate parameters (inertia weight w, cognitive coefficient C_1, and social coefficient C_2) across a range of 0.1 to 0.9.

1. Objective function (fitness): The common objective function minimized by both GA and PSO is the within-cluster sum of squares (WCSS) or sum of squared errors (SSE), which quantifies the internal cohesion of the clusters. A lower SSE indicates a superior, denser cluster arrangement.
2. Final K-means clustering: The K-means algorithm is run last, utilizing the optimal gBest centroid coordinates identified by the preceding GAPSO optimization as its starting points.

The Euclidean distance used in the K-means process is calculated as:

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

To ensure full reproducibility of the proposed framework, the hyperparameter settings for the hybrid optimization are rigorously defined based on empirical tuning and established literature, as summarized in Table 3. The GA phase utilizes a population size of 100 individuals to maintain genetic diversity and prevent premature convergence. A crossover rate (P_c) of 0.7 is applied to maximize genetic recombination, while the mutation rate (P_m) is dynamically set at $1/(10 \times len(bounds))$ to maintain stability across different problem dimensions. In the subsequent refinement phase, the PSO algorithm utilizes an inertia weight w of 0.7 and cognitive/social coefficients (C_1, C_2) of 1.5 to perform intensive local exploitation. The optimization process terminates after 100 generations or when the SSE reaches a stable convergence point.

Table 3 Detailed hyperparameter settings for GAPSO framework.

Phase	Parameter	Value/Method
Genetic Algorithm	Population Size (N_{pop})	10
	Max Generations	100
	Selection Method	Tournament selection
	Crossover Rate (P_c)	0.7
	Mutation Rate (P_m)	$1/(10 \times len(bounds))$
PSO Refinement	Inertia Weight (w)	0.7
	Acceleration (C_1, C_2)	1.5
	Swarm Size	100 (injected from GA)

To ensure complete methodological clarity and facilitate reproducibility, the step-by-step execution of the proposed Hybrid K-means+GAPSO framework is explicitly outlined in Pseudocode 1. The optimization process is systematically structured into three sequential phases. First, the GA conducts a global exploration across the solution space to identify promising regions and avoid local optima. Second, the optimal candidate solution from the GA phase is injected into the PSO algorithm to initialize an intensive local refinement. Finally, the refined centroids produced by PSO are utilized as the definitive starting points for the standard K-means algorithm to achieve optimal and stable convergence.

The mutation rate parameter, defined as the mathematical probability of random gene alterations occurring within an offspring's chromosome, was deliberately established at 0.1. The implementation of this specific probability threshold guarantees the continuous introduction of essential genetic diversity to prevent premature convergence without disrupting the stability of the optimal search trajectory.

Pseudocode 1 Hybrid GA-PSO for K-means centroid initialization.

```
// PHASE 1: GENETIC ALGORITHM (GLOBAL EXPLORATION)
1:  Initialize GA population P with N_pop individuals (randomly generate K centroids
per individual)
2:  FOR gen = 1 TO MaxGen DO
3:      FOR each individual i in P DO
4:          Evaluate fitness(i) using sum of squared errors (SSE)
5:      END FOR
6:      Select parents from P using tournament selection
7:      Apply Crossover with probability P_c to generate offspring
8:      Apply Mutation with probability P_m to offspring
9:      Update population P with new offspring
10: END FOR
11: Identify the best individual in P (minimum SSE) -> best_GA
```

```
// PHASE 2: PARTICLE SWARM OPTIMIZATION (LOCAL REFINEMENT)
12: Initialize PSO swarm S with N_pop particles distributed around best_GA
13: Set initial global best (gbest) = best_GA
14: FOR iter = 1 TO MaxIter DO
15:     FOR each particle p in S DO
16:         Evaluate fitness(p) using SSE
17:         IF fitness(p) < fitness(pbest_p) THEN
18:             Update personal best: pbest_p = p
19:         END IF
20:         IF fitness(p) < fitness(gbest) THEN
21:             Update global best: gbest = p
22:         END IF
23:     END FOR
24:     FOR each particle p in S DO
25:         Update velocity: v = w * v + c_1 * r_1 * (pbest_p - p) + c_2 * r_2 * (gbest -
p)
26:         Update position: p = p + v
27:     END FOR
28: END FOR
// PHASE 3: FINAL K-MEANS CLUSTERING
29: Initialize K-means starting centroids using final gbest
30: RUN standard K-means algorithm on dataset X until convergence
31: RETURN final cluster assignments and evaluation metrics
```

Furthermore, to provide a clear visual representation of the overall optimization methodology, the complete workflow of the proposed framework is illustrated in Figure 1. As depicted in the flowchart, the process begins with data preprocessing to handle multi-dimensional scaling and outliers. This is immediately followed by the global exploration phase using the GA to discover the most promising regions for centroid placement. The best candidate solution from the GA is then seamlessly injected into the PSO phase for intensive local refinement. Finally, the optimized centroids (gbest) are utilized by the standard K-means algorithm to accurately partition the dataset, concluding with a robust internal evaluation.

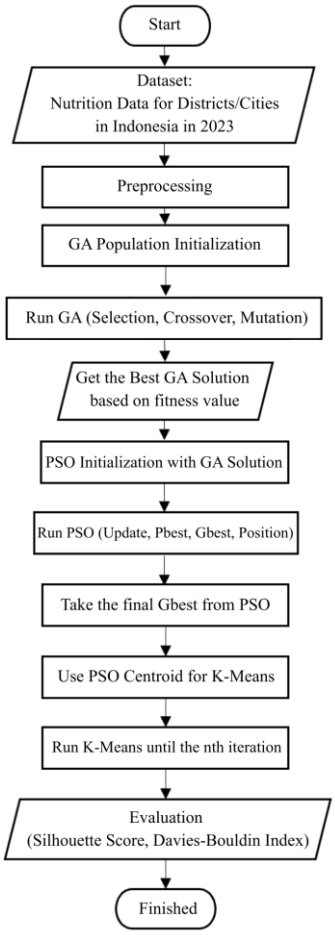

Figure 1 Flowchart of the hybrid approach.

3.4 Experimental Design and Evaluation

The model's performance was quantitatively assessed through a comparative study:

1. Model comparison: the Hybrid K-means + GAPSO model was compared against standard K-means and K-means + GA.
2. Parameter variation: all models were tested by varying the number of clusters (n) from 2 to 10.
3. Evaluation metrics: since the nutrition data was unlabeled, internal validation metrics were utilized to measure the clustering quality.

silhouette score (SC): measures the balance between cohesion (similarity within cluster, a(i)) and separation (distance from nearest cluster, b(i)). The goal was a higher score (closer to 1).

$$s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \quad (2)$$

Davies-Bouldin index (DBI): measures the average similarity between each cluster and its closest neighbor. The goal is a lower DBI value (closer to 0), indicating clusters are internally dense and externally well-separated.

$$DBI = \frac{1}{n}\sum_{i=1}^{n} max_{j\neq i}\left(\frac{S_i+S_j}{d(c_i,c_j)}\right) \quad (3)$$

Where S_i is the intra-cluster density of cluster i, and $d(c_i, c_j)$ is the distance between centroids c_i and c_j.

The execution of the refinement probability spectrum sequentially from 0.1 to 0.9 fundamentally functioned as a systematic sensitivity analysis. This evaluation specifically aimed to precisely measure the impact of PSO exploitation intervention on the GA exploration phase. The adoption of other static hyperparameters, such as the 0.1 mutation rate and the predefined population size, was strictly grounded in the empirical consensus within the metaheuristic optimization literature to maintain an optimal balance between search diversity and computational time efficiency.

To address the inherent stochasticity of metaheuristic optimization, the proposed framework implements a controlled execution protocol to ensure computational reliability. The standard K-means component is configured with $\text{n\_init} = 1$ using the specific centroids derived from the GAPSO phase. This configuration ensures that the final clustering performance is strictly dependent on the quality of the optimized initialization rather than multiple random restarts within the K-means algorithm itself. Furthermore, while GA and PSO are probabilistic by nature, a fixed random_state is utilized during the K-means execution to maintain reproducibility of the final partitioning stage. This approach allows for a direct assessment of how a single, intensive execution of global-local optimization can effectively overcome the local optima traps typically found in national nutritional datasets.

The algorithmic performance evaluation in this study strictly avoids single-run execution, considering the inherent stochastic nature of both GA and PSO. To ensure experimental robustness, the testing was systematically executed through 10 independent runs for every hyperparameter configuration. The initialization of the random seed was dynamically randomized at each iteration to capture the true

probabilistic behavior and prove the model's reliability against variance fluctuations. The reporting of computational results utilizes mean and standard deviation (Std) values to present a highly accurate and academically rigorous convergence overview.

4 Results and Discussion

This section presents the comparative performance analysis of the clustering models, followed by a discussion interpreting these findings within the context of the research objectives. The primary goal was to empirically determine if the proposed Hybrid K-means+GAPSO framework provides a more robust and accurate centroid initialization compared to the standard K-means and the single-metaheuristic K-means+GA baseline. Furthermore, this section systematically details the determination of the optimal cluster count (K) and translates the resulting regional nutritional typologies into practical policy insights.

4.1 Clustering Test Results with GA-PSO

Table 4 presents the summarized results of the optimal K-means+GAPSO configurations for each cluster size across 10 independent runs. The evaluation metrics specifically display the highest average performance accompanied by their respective standard deviations. The documentation of these minimal variance values empirically signifies the robust convergence capability of the proposed hybrid architecture across multiple randomized executions.

Table 4 Summary of optimal K-means+GAPSO configurations across 10 independent runs.

Cluster Size (K)	Optimal Refinement Probability	Silhouette Score (Mean ± Std)	Davies-Bouldin Index (Mean ± Std)
2	0.1	0.7450 ± 0.0000	0.5119 ± 0.0000
3	**0.3**	**0.7130 ± 0.0036**	**0.4708 ± 0.0178**
4	0.7	0.6520 ± 0.0336	0.5261 ± 0.0052
5	0.3	0.5924 ± 0.0114	0.5247 ± 0.0156
6	0.2	0.5720 ± 0.0071	0.5207 ± 0.0138
7	0.3	0.5594 ± 0.0038	0.5560 ± 0.0108
8	0.8	0.5482 ± 0.0060	0.5683 ± 0.0175
9	0.4	0.5336 ± 0.0121	0.5809 ± 0.0129
10	0.9	0.5279 ± 0.0072	0.5909 ± 0.0079

4.2 Determination of Optimal Cluster Count (K)

To objectively determine the optimal number of clusters (K), this study conducted a comprehensive exploration across a range of K = 2 to K = 10. Visually subjective heuristic approaches, such as the traditional elbow method, were deliberately bypassed because they often yield ambiguous inflection points in

complex, high-dimensional datasets. Instead, the optimal cluster count was determined by focusing on direct, quantitative measures of structural quality: the silhouette score and the Davies-Bouldin index (DBI). By plotting these metrics against the varying K values, a clear 'sweet spot' was mathematically identified. Specifically for the proposed Hybrid K-means+GAPSO framework, K = 3 emerged as the definitive optimal configuration, successfully achieving a dual-objective peak by recording higher silhouette scores (indicating superior cohesion and separation) and lower DBIs (indicating maximum intra-cluster compactness) compared to alternative counts such as K = 4 or K = 5.

4.3 Clustering Performance Comparison

The evaluation of the clustering performance encompassed three distinct models: standard K-means (as the baseline), K-means with GA initialization (K-means+GA), and the proposed hybrid framework (K-means+GAPSO).

The execution of these models on the National Nutrition Dataset spanned a continuous range of clusters (K = 2 to K = 10) across 10 independent runs. Quantitative assessment of cluster quality utilized two standard internal validation metrics: the silhouette score for measuring inter-cluster separation (higher is better) and the Davies-Bouldin index for measuring intra-cluster compactness relative to inter-cluster separation (lower is better). The performance trends and comparative results are visually illustrated in Figures 2 and 3.

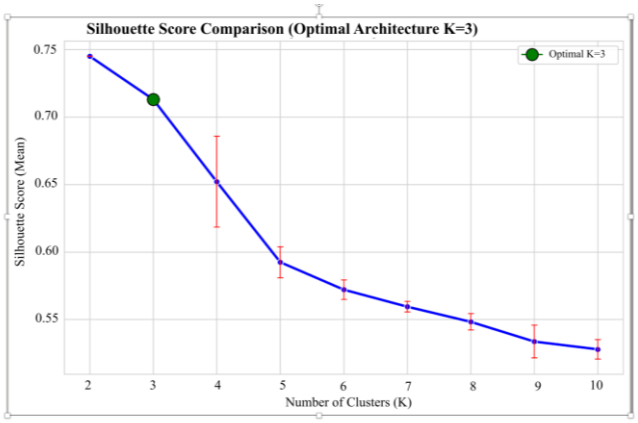

Figure 2 Comparison of silhouette score trends (K = 2 to K = 10).

Figure 2 illustrates the comparative performance trend as the number of clusters (K) sequentially increases from 2 to 10. The visual trajectory in the graph confirms a general pattern of performance degradation across all models,

reflecting a typical phenomenon when data partitioning becomes excessively granular. The standard K-means algorithm exhibits the most rapid and severe metric decline, underscoring its inherent vulnerability to local optima. The proposed K-means+GAPSO architecture, in sharp contrast, consistently maintains its performance levels in the upper tier of the evaluation. This computational stability is best quantified at the optimal point K = 3, where the K-means+GAPSO configuration with a 0.3 refinement probability achieves a highly stable silhouette mean of 0.7130 (±0.0036). This metric achievement represents a substantial improvement in cluster separation compared to the standard K-means baseline. This sustained algorithmic superiority confirms that the sequential GAPSO initialization mechanism provides a robust methodological shield against the instability typically plaguing distance-based clustering in complex data. The resulting data partitioning consequently delivers reliable cluster separation for nuanced public health analysis.

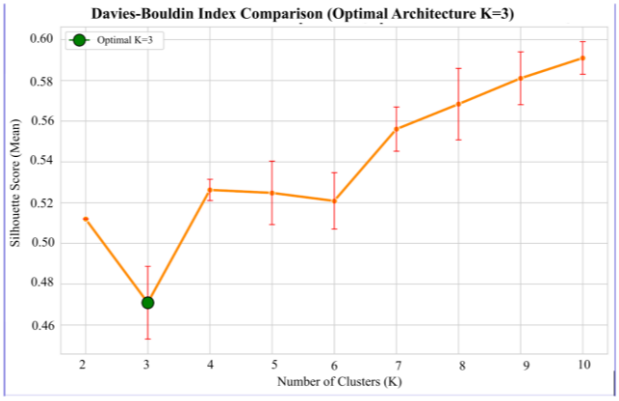

Figure 3 Comparison of Davies-Bouldin index Trends (K = 2 to K = 10).

Figure 3 illustrates the comparative performance of all evaluated algorithms concerning the Davies-Bouldin index where lower values signify superior cluster compactness and separation. The visual representation vividly confirms the instability of the baseline method, as the standard K-means algorithm consistently occupies the upper region, exhibiting the highest (worst) DBI scores across most of the K range. The proposed hybrid approaches, particularly the K-means+GAPSO variants, in sharp contrast, successfully minimized this metric, clustering tightly near the graph's lower bound. The profound impact of this optimization is best demonstrated at the optimal cluster count K = 3. The most superior variant, K-means+GAPSO with a 0.3 refinement probability, achieved the lowest highly stable DBI mean of 0.4708 (±0.0178). This statistical achievement represents a substantial reduction of the DBI score compared to the

standard K-means baseline, confirming the framework's exceptional ability to form highly dense and well-separated clusters. This crucial reduction in DBI significantly boosts the methodological confidence in the results, making the derived nutritional classifications highly reliable for evidence-based policy interventions.

The synthesis of the empirical evidence from both Figure 2 and Figure 3 unequivocally establishes the K-means+GAPSO (prob = 0.3) variant as the superior algorithm for this study. The 0.3 probability setting, while several hybrid configurations demonstrated competence over the baseline, achieved the rare dual-objective optimization at the optimal K = 3, namely, maximizing cluster separation (highest silhouette score) while simultaneously enforcing the strictest cluster compactness (lowest DBI). This distinct performance peak identifies the K-means+GAPSO 0.3 configuration not merely as a statistical improvement but as the most structurally valid model for this specific nutritional dataset. The implementation of this optimal model offers the highest degree of confidence for ensuring that the resulting policy recommendations are based on solid, well-defined data groupings.

Visual evaluation of the spatial distribution and structural boundaries of the formed clusters within the high-dimensional nutritional dataset utilized a dimensionality reduction technique via 3D principal component analysis (PCA) strictly for post-clustering visualization. The projection of the multi-dimensional feature space onto three primary principal components, as illustrated in Figures 4, 5, and 6, compares the structural integrity of the clusters generated by Pure K-means, K-means+GA, and the optimal configuration of the Hybrid K-means+GAPSO framework (refinement probability = 0.3).

The visual progression across the three plots definitively corroborates the quantitative performance metrics. The Pure K-means visualization (Figure 4) displays diffuse cluster boundaries with noticeable spatial overlap, reflecting the algorithm's susceptibility to local optima in highly skewed data. The K-means+GA model (Figure 5) shows marginal structural improvements. The Hybrid K-means+GAPSO model (Figure 6), however, demonstrates vastly superior structural coherence. The optimal K = 3 clusters exhibit highly distinct spatial separation and strong internal cohesion, with boundaries that are tightly packed and clearly defined. This visual validation confirms the efficacy of the Hybrid GAPSO initialization in optimizing the centroids and successfully navigating complex nutritional data topologies to isolate distinct, policy-relevant regional typologies with minimal structural overlap.

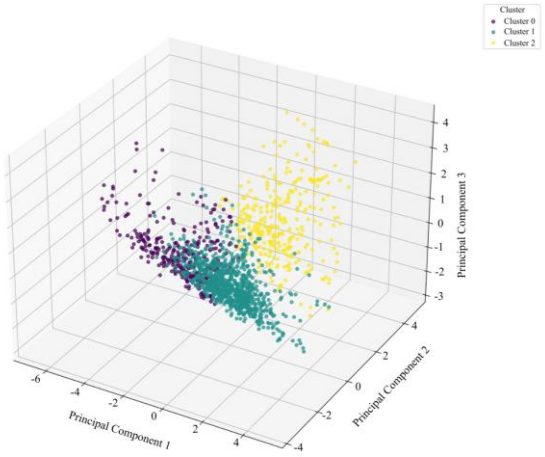

Figure 4 PCA visualization: Pure K-means (K = 3).

Figure 4 illustrates the 3D PCA projection of the spatial distribution generated by the baseline Pure K-means algorithm. The visualization reveals diffuse cluster boundaries with noticeable spatial overlap among the groups, particularly between Clusters 0, 1, and 2. This structural blending visually highlights the limitations of standard K-means in cleanly separating highly skewed, high-dimensional data without metaheuristic optimization.

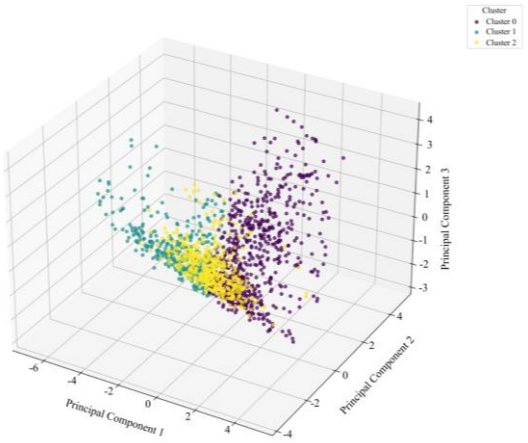

Figure 5 PCA visualization: K-means + GA (K = 3).

Figure 5 depicts the spatial distribution achieved by incorporating the genetic algorithm (K-means+GA). The global search mechanism marginally improves the cluster boundaries and reduces extreme dispersion compared to the baseline, yet significant overlapping remains evident in the dense central region. This

visual evidence indicates the struggle of relying solely on GA without intensive local refinement to fully disentangle the complex data topologies.

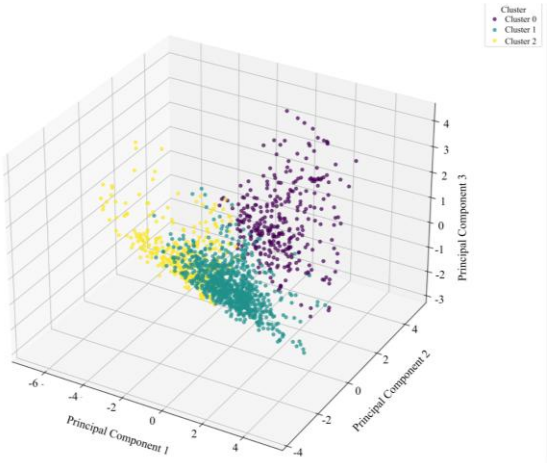

Figure 6 PCA visualization: K-means + GAPSO (K = 3, prob = 0.3).

Figure 6 illustrates the optimal spatial structure achieved by the proposed Hybrid K-means+GAPSO framework. The clusters, in stark contrast to the previous baseline models, exhibit vastly superior structural coherence, characterized by tightly packed internal distributions and distinctly delineated boundaries. This clear visual separation directly corroborates the high silhouette score, confirming the success of the synergistic combination of global exploration and local exploitation in isolating distinct nutritional typologies with minimal structural overlap.

4.4 Nutritional Cluster Profiling

To translate the statistical clustering results into actionable policy insight, we analyzed the geographical and nutritional characteristics associated with the three identified clusters. This profiling indicated that the K-means+GAPSO model effectively grouped regions with shared socioeconomic and nutritional patterns within the national dataset.

Of particular significance to the Makan Bergizi Gratis (MBG) initiative is Cluster 1, which groups high-risk regions characterized by severe nutritional vulnerability specifically, a high PoU and low protein intake. Our analysis revealed a critical concentration of these regions in Eastern Indonesia, particularly within Papua. The algorithm grouped districts such as Kabupaten Nduga, Kabupaten Jayawijaya, and Kabupaten Pegunungan Arfak into this

cluster. The identification of these specific districts highlights a strong association between this cluster's profile and the potential need for targeted nutritional aid and infrastructure support to help address malnutrition in these geographically challenging areas.

In contrast, Cluster 0 encompasses primarily highly urbanized strata, frequently located in Java and major economic centers. Representative samples such as Kota Tangerang (Banten), Kabupaten Sumedang (Jawa Barat), and Kabupaten Lebak (Banten) fall into this category. These regions exhibit high caloric and protein consumption patterns, which are typically associated with better market access and purchasing power. For regions within this demographic profile, policy discussions might better be directed towards nutritional quality education and lifestyle-related health interventions rather than emergency basic food security.

Meanwhile, Cluster 2 emerged as the most dominant group, encompassing regions with standard nutritional profiles hovering near the national average. Geographically widespread, this cluster includes regencies such as Kabupaten Karo (Sumatera Utara), Kabupaten Gowa (Sulawesi Selatan), and Kabupaten Kolaka Utara (Sulawesi Tenggara). The profile of these areas suggests relatively stable food systems that may require maintenance strategies rather than emergency relief.

The fundamental nature of this spatial clustering framework is inherently exploratory. The resulting data partitions effectively highlight associative typologies of nutritional vulnerability across regions without implying any direct causal relationships among the underlying socioeconomic variables. The translation of these computational clustering results into actionable government policies, such as the Makan Bergizi Gratis program, must therefore proceed with analytical caution. The formulation of definitive public health interventions requires complementary causal investigations and comprehensive on-the-ground validations to confirm the contextual dynamics of the targeted regions.

5 Conclusion

This study successfully mitigated the inherent instability of the standard K-means algorithm stemming from its sensitivity to random centroid initialization. The development of the sequential K-means+GAPSO framework synergizes the broad global search capabilities of genetic algorithms with the precise local refinement of particle swarm optimization. Empirical validation on the Indonesian National Nutrition Dataset (2021–2023) across 10 independent runs identified $K = 3$ as the optimal cluster count for regional nutritional mapping. The comparative analysis definitively positioned the K-means+GAPSO configuration with a 0.3 refinement probability as the superior model. This hybrid architecture

achieved a dual-objective optimization, recording a highly stable silhouette mean of 0.7130 (±0.0036) and a Davies-Bouldin index mean of 0.4708 (±0.0178). These evaluation metrics revealed substantial performance improvements in cluster separation and intra-cluster compactness compared to the standard K-means baseline. This research provides a significant methodological contribution by delivering a robust and highly stable clustering tool that transcends technical improvements to offer practical value. The precise data partitioning generated by this model effectively maps regional nutritional typologies, even though its exploratory nature highlights associative spatial patterns rather than direct causal relationships. The identification of regions sharing similar socioeconomic and nutritional vulnerabilities assists policymakers in targeting areas more accurately. This advanced framework ultimately empowers the government to formulate targeted, evidence-based public health interventions, such as the Makan Bergizi Gratis (Free Nutritious Meal) program, with a significantly higher degree of analytical confidence and strategic accuracy.

Acknowledgement

This research was funded by the Faculty of Engineering, Universitas Diponegoro, Indonesia.

References

[1] Tsipi, L., Michailidis, E.T., Karavolos, M., Vouyioukas, D. & Kanatas, A.G., *AI-driven Aerial Relay Placement and Power Allocation for Enhanced Secrecy in Untrusted UAV-enabled Networks*, IEEE Access., **13**, pp. 99970-99984, 2025. doi:10.1109/ACCESS.2025.3577053.

[2] Yin, X., Rong, Y., Li, L., He, W., Lv, M. & Sun, S., *Health State Prediction Method Based on Multi-featured Parameter Information Fusion*, Appl. Sci., **14**(15), p. 6809, Aug. 2024. doi:10.3390/app14156809.

[3] Latief, M.A., Pandiya, R. & Putri, A.L.R., *Least Square-based Differential Evolution Algorithm for N-dimensional Data Clustering Problem*, Eurasian J. Math. Comput. Appl., 2025. doi:10.32523/2306-6172-2025-13-3-36-49.

[4] Abdo, A., Abdelkader, O. & Abdel-Hamid, L., *SA-PSO-GK++: A New Hybrid Clustering Approach for Analyzing Medical Data*, IEEE Access., **12**, pp. 12501-12516, 2024. doi:10.1109/ACCESS.2024.3350442.

[5] Latief, M.A., Candraningtyas, R.G., Khomsah, S. & Pandiya, R., *Grey Wolf Algorithm for Optimizing K-modes in Determining Initial Centroid*, in 2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), Mataram, Indonesia: IEEE, pp. 429-436, Nov. 2024. doi:10.1109/COMNETSAT63286.2024.10862708.

[6] Suwirmayanti, N.L.G.P., Gede Darma Putra, I.K., Sudarma, M., Sukarsa, I.M. & Setyaningsih, E., *IWOKM-GA Hybrid Method to Improve Clustering Accuracy in Banking Data*, Int. J. Comput. Digit. Syst., **18**(1), pp. 1-11, Apr. 2025. doi:10.12785/ijcds/1571110934.

[7] Karishma & Kumar, H., *A Novel Hybrid Model for Task Scheduling based on Particle Swarm Optimization and Genetic Algorithms*, Math. Eng., **6**(4), pp. 559-606, 2024. doi:10.3934/mine.2024023.

[8] Politi, R.R. & Tanyel, S., *Minimizing Delay at Closely Spaced Signalized Intersections Through Green Time Ratio Optimization: A Hybrid Approach with K-means Clustering and Genetic Algorithms*, IEEE Access., **13**, pp. 43981-43999, 2025. doi:10.1109/ACCESS.2025.3549970.

[9] Zhou, Y. & Yan, *The Application of Big Data Technology Combined with Improved Genetic Simulated Annealing Algorithm in the Construction of Intelligent Delivery Models*, Int. J. Mechatron. Appl. Mech., **1**(20), Jun. 2025. doi:10.17683/ijomam/issue20.7.

[10] Hassan, E. et al., *A Hybrid K-Means++ and Particle Swarm Optimization Approach for Enhanced Document Clustering*, IEEE Access., **13**, pp. 48818-48840, 2025. doi:10.1109/ACCESS.2025.3535226.

[11] Pramudya, R.I., Kurniawan, T.A., Candra, M.H., Onn, C.W. & Dissanayake, K.K., *Advancing Unsupervised Clustering: A Systematic Review of Hybrid K-means and Metaheuristic Optimization Algorithms in Data Mining*, Comput. Sci. Rev., **61**, 100972, Aug. 2026. doi: 10.1016/j.cosrev.2026.100972.

[12] Bafadal, I.Z., Ardi, R. & Yuraisyah Salsabila, N., *A Proposed Reverse Logistics Network for End-of-life Electric Vehicle Battery Management in the Jakarta Greater Area: A MILP Approach*, World Electr. Veh. J., **16**(8), 2025. doi:10.3390/wevj16080476.

[13] Li, K., Wang, Z., Zhou, Y. & Li, S., *Lung Adenocarcinoma Identification Based on Hybrid Feature Selections and Attentional Convolutional Neural Networks*, Math. Biosci. Eng., **21**(2), pp. 2991-3015, 2024. doi:10.3934/mbe.2024133.

[14] Zaghdoud, R., Rhaiem, O.B., Amara, M., Mesghouni, K. & Galet, S., *A Hybrid Genetic Algorithm Approach based on Patient Classification to Optimize Home Health Care Scheduling and Routing*, Eng. Technol. Appl. Sci. Res., **14**(4), pp. 15099-15105, Aug. 2024. doi:10.48084/etasr.7649.

[15] Seo, Y., Lim, D., Lee, J., So, J. & Kim, H., *A Framework for Evaluating Driving Patterns of Automated Vehicle Based on On-Road Driving Records*, IEEE Access., **13**, pp. 67894-67907, 2025. doi:10.1109/ACCESS.2025.3556299.

[16] Wang, X.L., Guo, H., Wen, F. & Wang, K., *Optimized low carbon scheduling strategy of integrated energy sources considering load aggregators*, AIP Adv., **14**(4), 045228, Apr. 2024. doi:10.1063/5.0190409.
[17] Li, J., Ma, Y., Li, H., Liu, Y. & Li, Y., *A Load Classification Method Based on Hybrid Clustering of Continuous–Discrete Electricity Consumption Characteristics*, Processes, **13**(4), 1208, Apr. 2025. doi:10.3390/pr13041208.
[18] National Food Agency, *Food Open Data: Data for 2021–2023,* 2024. Available: https://data.badanpangan.go.id/. (15 June 2026)