

685 ~ 699 - ID 889.pdf

by Turnitin Web

Submission date: 14-Jul-2026 12:59PM (UTC+0800)

Submission ID: 2997825787

File name: 685_699_-_ID_889.pdf (945.72K)

Word count: 5142

Character count: 31712

METAHEURISTIC-ENHANCED HIERARCHICAL AGGLOMERATIVE CLUSTERING FOR NON-COMMUNICABLE DISEASE RISK FACTOR PROFILING IN BANTEN PROVINCE

Tb Ai Munandar^{1*}, Muhammad Fairuzabadi², and Herison Surbakti³

¹Informatics, Universitas Bhayangkara Jakarta Raya, Indonesia

²Informatics Eng. Dept., Universitas PGRI Yogyakarta, Indonesia

³Information and Communication Technology, International College, Rangsit University, Thailand

tbaumunandar@gmail.com¹, fairuz@upy.ac.id², herison.s@rsu.ac.th³

*Corresponding Author

Received June 30, 2026; Revised July 11, 2026; Accepted July 13, 2026; Published July 31, 2026

ABSTRACT

Non-communicable diseases (NCDs) account for approximately 74% of all global deaths annually and impose a disproportionate burden on developing nations, including Indonesia. In Banten Province, epidemiological surveillance data reveal highly heterogeneous distributions of NCD behavioral risk factors across 149 sub-districts (aggregated from facility-level surveillance records), complicating the design of targeted public health interventions. This study proposes and evaluates four metaheuristic-enhanced Hierarchical Agglomerative Clustering (HAC) frameworks—GA-HAC, PSO-HAC, GWO-HAC, and WOA-HAC—applied to five behavioral risk indicators: excessive fat/oil consumption (efc), excessive salt consumption (esc), excessive sugar consumption (esuc), tobacco smoke exposure (etcs), and active smoking (smoke). Each metaheuristic optimizes a composite fitness function combining the Silhouette Score and the Davies-Bouldin Index on log1p-transformed features. Across 20 independent runs, WOA-HAC achieves the best mean Silhouette Score (0.4745) and lowest mean Davies-Bouldin Index (0.6795), significantly outperforming PSO-HAC and GWO-HAC on both criteria (Kruskal-Wallis, $p < 0.001$), though no single algorithm dominates the Calinski-Harabasz criterion. The resulting three-tier solution reveals a Low Risk tier ($n=25$, 16.8%), a smoking-predominant tier concentrated in Kabupaten Pandeglang ($n=26$, 17.4%), and a Very High Risk tier spanning all eight administrative units ($n=98$, 65.8%) with majority-level prevalence across all five indicators simultaneously. Elevated NCD risk in Banten Province thus appears widespread rather than confined to a small urban enclave, warranting province-wide, multi-component prevention programming.

Keywords: Non-communicable disease, hierarchical agglomerative clustering, genetic algorithm, particle swarm optimization, metaheuristic optimization.

15

INTRODUCTION

Non-communicable diseases (NCDs) are the world's leading cause of death, responsible for approximately 41 million fatalities annually—equivalent to 74% of all global deaths (World Health Organization, 2025). Cardiovascular disease, cancer, chronic respiratory illness, and diabetes together account for over 80% of this burden (Rusyda, 2025). NCDs disproportionately affect low-and middle-income countries, where healthcare systems are least equipped to manage long-term chronic conditions and where prevention resources remain inadequately deployed. The human and economic costs are staggering: the World Health Organization projects that NCDs will cost the global

economy more than USD 30 trillion by 2030 in lost productivity and healthcare expenditure, with the sharpest losses concentrated in South and Southeast Asia.\

Indonesia exemplifies the dual burden facing rapidly developing economies. National health survey data (Riskesdas 2023) (Badan Penelitian dan Pengembangan Kesehatan, 2023) document an accelerating NCD trajectory: hypertension affects 28.5% of adults, diabetes mellitus 10.9%, and stroke 10.9 per 1,000 individuals—all substantially above 2018 estimates. These aggregate figures mask profound geographic heterogeneity: risk factor prevalence varies dramatically between provinces and, critically, between sub-districts within provinces. Urbanization, the nutrition transition toward energy-dense processed diets, rising tobacco use, and physical inactivity collectively drive this epidemic, yet their relative contributions vary enormously by locality. In Banten Province specifically, this sub-district-level heterogeneity has direct planning consequences: provincial and district health offices currently allocate NCD prevention budgets and staff largely on the basis of aggregate district-level prevalence, a resolution too coarse to distinguish a handful of extremely high-burden sub-districts from their lower-risk neighbors within the same district. Without sub-district-resolution evidence, high-need communities can remain under-served for years while lower-risk areas receive comparable resources, directly undermining the efficiency and equity of provincial NCD prevention planning.

Banten Province—a densely populated region of 12 million residents on Java's western tip—presents a particularly compelling setting for sub-district-level analysis. Its eight administrative units span a socioeconomic spectrum from the highly urbanized, Jakarta-adjacent Kota Tangerang Selatan to the predominantly rural Lebak and Pandeglang Districts. Provincial health data (Dinas Kesehatan Provinsi Banten, 2022) confirm that NCD risk factor prevalence is not uniformly distributed: some sub-districts report prevalence levels at or above the majority, across multiple simultaneous risk behaviors, while others exhibit relatively low exposure profiles. This heterogeneity implies that province-wide or district-wide health interventions—designed around average risk profiles—will systematically under-serve the highest-burden communities while over-allocating resources to lower-risk areas.

Data-driven risk stratification through clustering analysis offers a principled solution. By grouping sub-districts with similar risk factor profiles, clustering reveals natural tiers that can anchor differentiated intervention strategies. Hierarchical Agglomerative Clustering (HAC) is particularly well suited to this task: it produces an interpretable dendrogram without requiring the number of clusters to be pre-specified, and Ward's linkage criterion generates compact, well-separated clusters appropriate for risk tier identification (Rokach, 2009; Ward, 1963). However, standard HAC does not optimize any global quality criterion—it greedily merges pairs based on a local linkage rule—leaving room for improvement through global optimization. Because the five behavioral risk indicators considered here frequently co-occur and reinforce one another within a community, a single-indicator ranking (e.g., sorting sub-districts by smoking prevalence alone) cannot capture the joint, multidimensional risk profile that actually

determines intervention need. Clustering analysis is appropriate precisely because it partitions sub-districts on the basis of their combined similarity across all five indicators simultaneously, revealing coherent multi-behavioral risk tiers rather than a ranking along any single axis.

Metaheuristic algorithms provide a powerful mechanism for this optimization. Genetic algorithms (GA) search discrete hyperparameter spaces through population-based evolutionary operations (Caparrini & Arroyo, 2026; Hameed & Alkhzaimi, 2025). Particle Swarm Optimization (PSO) leverages collective swarm intelligence to efficiently perform continuous optimization (Fan et al., 2026; Triantali & Parsopoulos, 2025). Grey Wolf Optimizer (GWO) and Whale Optimization Algorithm (WOA) draw on biologically inspired search dynamics that balance exploration and exploitation (Kyrou et al., 2026; Yousef et al., 2026; Zhou & Hao, 2025). Applying these algorithms to optimize HAC configurations or cluster centroid positions can yield cluster structures that are more cohesive and better separated than those obtained by the greedy baseline procedure.

Despite growing interest in metaheuristic clustering, applications targeting sub-district-level NCD risk factor data in Indonesian provinces remain scarce. Prior work in this region has relied primarily on k-means, which is sensitive to initialization (Bombelli et al., 2026) and imposes spherical-cluster assumptions (Wani, 2024) that may not reflect the true distribution of risk factors. No published study has compared multiple metaheuristic-optimized HAC frameworks on real-world Indonesian NCD surveillance data using a unified evaluation design. This study addresses that gap. The scientific contribution of this work is therefore twofold: it is, to our knowledge, the first study to benchmark four distinct metaheuristic paradigms—evolutionary search (GA), swarm intelligence (PSO), and the bio-inspired leader-following and spiral-search mechanisms of GWO and WOA—against a common HAC-Ward baseline and a single composite fitness criterion on real Indonesian sub-district NCD surveillance data, and it translates the best-performing configuration into epidemiologically interpretable, sub-district-resolution risk tiers rather than stopping at algorithmic comparison alone. Specifically, this study aims to (i) benchmark GA-, PSO-, GWO-, and WOA-enhanced HAC against the HAC-Ward baseline using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score, and (ii) translate the best-performing clustering solution into actionable, sub-district-resolution NCD risk tiers for Banten Province's provincial health authorities.

The contributions of this work are: (1) a comprehensive NCD risk factor dataset covering all 149 sub-districts of Banten Province is compiled and analyzed; (2) a unified metaheuristic-HAC framework is proposed in which GA, PSO, GWO, and WOA each optimize a composite fitness function combining Silhouette Score and Davies-Bouldin Index; (3) all four variants are rigorously benchmarked against baseline HAC-Ward on three internal validation metrics; (4) the resulting clusters are epidemiologically interpreted, yielding actionable risk tier profiles at sub-district resolution; and (5) practical policy recommendations are derived for Banten Province's health authorities based on the identified Very High Risk sub-districts.

RESEARCH METHOD

The methodology comprises five sequential components: (A) data collection and pre-processing, (B) baseline HAC and cluster number selection, (C) composite fitness function design, (D) metaheuristic optimization (GA-HAC, PSO-HAC, GWO-HAC, WOA-HAC), and (E) cluster validation. These four algorithms were selected to span the principal search paradigms used in the clustering-optimization literature: GA represents discrete, population-based evolutionary search over HAC configuration choices; PSO represents continuous, velocity-based swarm intelligence; and GWO and WOA represent more recent bio-inspired algorithms driven by leader-following and spiral-search mechanisms, respectively. Comparing all four under an identical fitness function and validation protocol isolates the relative merits of discrete configuration search versus continuous centroid optimization from algorithm-specific effects.

Data Collection and Pre-processing

The dataset comprises behavioral NCD risk factor surveillance records from the Banten Provincial Health Office's integrated surveillance system, covering all eight administrative units of Banten Province: Kabupaten Tangerang, Lebak, Pandeglang, and Serang; Kota Tangerang, Tangerang Selatan, Cilegon, and Serang. Raw records are reported at the primary health center (Puskesmas) level; because multiple Puskesmas can serve the same sub-district, records were aggregated (summed) by sub-district prior to analysis, yielding a final analytic sample of 149 unique sub-districts. Five features are included: efc (excessive fat/oil consumption), esc (excessive salt consumption), esuc (excessive sugar consumption), etcs (secondhand tobacco smoke exposure), and smoke (active smoking prevalence), all expressed as sub-district-level counts.

Descriptive analysis revealed right-skewed distributions for all five features (see Table 1), with coefficients of variation exceeding 170% and median values well below corresponding means, consistent with the presence of high-risk outlier sub-districts within a predominantly lower-risk population. The aggregated dataset contained no missing values. Because the raw counts span several orders of magnitude (Table 1), a log1p transformation was applied to each feature prior to scaling, preventing one or two extreme outlier sub-districts from dominating Euclidean distance computations and forcing degenerate clustering solutions. All log-transformed features were then normalized to [0,1] via Min-Max scaling (Equation 1) to ensure equal contribution to distance computations. Prior to normalization, records were checked for internal consistency and range plausibility and cross-referenced against Dinas Kesehatan Provinsi Banten's published district-level summaries to confirm that sub-district-level aggregates were consistent with reported district totals.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Table 1. Descriptive Statistics Of NCD Risk Factors

Feature	Min	Max	Mean	Median	Std Dev
efc	0.00	4014.00	234.28	35.00	483.13
esc	0.00	4442.00	224.76	35.00	497.81
esuc	0.00	3891.00	213.62	34.00	461.77
etcs	0.00	7598.00	379.59	82.00	780.79
smoke	0.00	3254.00	246.59	62.00	422.33

Baseline HAC and Cluster Number Selection

HAC with Ward's minimum variance linkage [5] was applied to the normalized dataset as the baseline. Ward's criterion merges clusters to minimize the increase in total within-cluster variance at each step, as formalized in Equation 2. The dendrogram (Figure 1) visualizes the full merging history and was used to guide the selection of the number of clusters.

$$\Delta(C_i, C_j) = \frac{|C_i| \cdot |C_j|}{|C_i| + |C_j|} \cdot \|\mu_i - \mu_j\|^2 \quad (2)$$

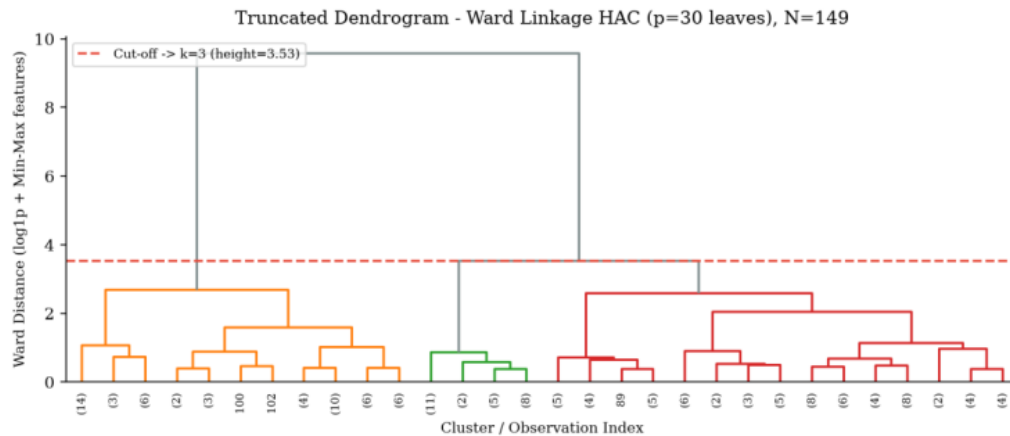


Figure 1. HAC-Ward dendrogram of 149 sub-districts (truncated to last 20 merges).

The optimal number of clusters, k , was determined by evaluating $k \in \{2, \dots, 7\}$ across three internal validation metrics: Silhouette Score (higher is better), Davies-Bouldin Index (lower is better), and Calinski-Harabasz Score (higher is better). Figure 2 shows the metric trajectories. The two-cluster solution achieves the best value on all three metrics simultaneously (Silhouette=0.5643, DB=0.6117, CH=288.83). Nevertheless, the three-cluster solution ($k=3$: Silhouette=0.3924, DB=0.7954, CH=222.09) was selected because it produces three epidemiologically interpretable risk tiers—Low Risk, a smoking-predominant tier, and Very High Risk—directly mapping to the three-tier prevention framework recommended by WHO and Indonesia's Ministry of Health, and because its smallest cluster (26 sub-districts, 17.4% of $n=149$) comfortably exceeds the 5% minimum cluster-size threshold used in the composite fitness function (Equation 3).

For $k \geq 4$, all three metrics deteriorate or fluctuate without a clear interpretive gain.

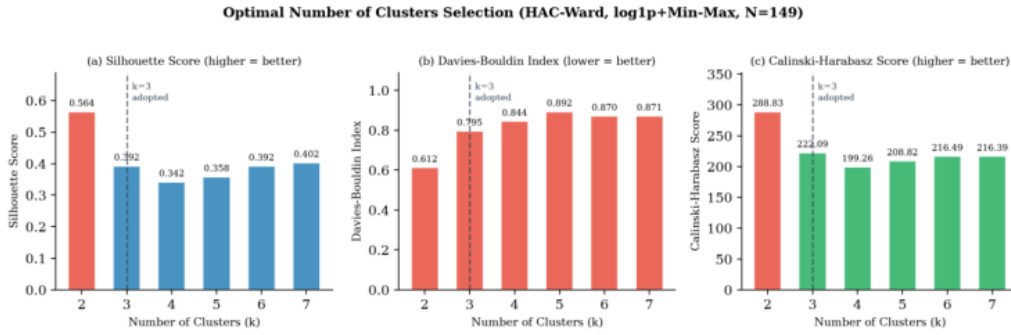


Figure 2. Internal validation metrics for $k=2$ to $k=7$ (HAC-Ward).

Composite Fitness Function

All four metaheuristics optimize the same composite fitness function (Equation 3), which combines Silhouette Score and the reciprocal of Davies-Bouldin Index. A minimum cluster size penalty p is applied if any cluster contains fewer than 5% of $n=149$ sub-districts (≈ 7 sub-districts), discouraging degenerate solutions with negligible clusters.

$$f = 0.6 \times s + 0.4 \times \frac{1}{1 + DB} - p \quad (3)$$

Weights of 0.6 (Silhouette) and 0.4 (DB term) reflect the greater prevalence of the Silhouette Score as the primary indicator in the clustering literature [14]. The penalty function returns $f = -1$ when the size constraint is violated, removing infeasible solutions from selection.

Metaheuristic Optimization

GA-HAC formulates the linkage method (ward/complete/average/single) and distance metric (euclidean/manhattan/cosine) as a discrete chromosome [linkage_idx, metric_idx], yielding a $4 \times 3 = 12$ -point search space. Parameters: population $P=30$, crossover rate 0.8, mutation rate 0.05, 50 generations, tournament selection (size=3).

PSO-HAC, GWO-HAC, and WOA-HAC optimize the positions of $k=3$ centroids in the 5-dimensional normalized feature space. Each agent encodes a candidate centroid set $C = \{c_1, c_2, c_3\}$, $c_i \in [0,1]^5$. All three algorithms use 25 agents, 70 iterations, and initialize populations near Ward-derived centroids with Gaussian perturbation ($\sigma=0.03$) to ensure feasible starting solutions. PSO velocity and position updates follow Equations 4-5 ($w=0.5$, $c_1=c_2=1.8$):

$$v_i^{(t+1)} = w \cdot v_i^t + c_1 \cdot r_1 \cdot (pbest_i - x_i^t) + c_2 \cdot r_2 \cdot (gbest - x_i^t) \quad (4)$$

$$x_i^{(t+1)} = x_i^t + v_i^{(t+1)} \quad (5)$$

GWO updates positions via the alpha-beta-delta hunting hierarchy (Equation 6):

$$\begin{aligned} D_{\alpha} &= |C_1 \cdot X_{\alpha} - X| \\ X_1 &= X_{\alpha} - A_1 \cdot D_{\alpha} \\ X^{(t+1)} &= \frac{X_1 + X_2 + X_3}{3} \end{aligned} \quad (6)$$

WOA alternates with equal probability ($p=0.5$) between encircling (Equation 7) and logarithmic spiral (Equation 8) mechanisms:

$$\begin{aligned} D &= |C \cdot X^*(t) - X(t)| \\ X^{(t+1)} &= X^*(t) - A \cdot D \end{aligned} \quad (7)$$

$$X^{(t+1)} = D' \cdot e^{bl} \cdot \cos(2\pi l) + X^*(t) \quad (8)$$

Cluster Validation Metrics

Clustering quality is assessed using three complementary metrics: silhouette score, Davies-Bouldin Index, and Calinski-Harabasz Score.

$$X^{(t+1)} = D' \cdot e^{bl} \cdot \cos(2\pi l) + X^*(t) \quad (9)$$

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \frac{\sigma_i + \sigma_j}{d(\mu_i, \mu_j)} \quad (10)$$

$$CH_{score} = \left[\frac{Trace B_k}{Trace W_k} \right] x \left[\frac{n-k}{k-1} \right] \quad (11)$$

The Calinski-Harabasz Score measures the ratio of between-cluster to within-cluster dispersion. Higher values indicate more compact, better-separated clusters. Together, these three metrics provide a comprehensive, multi-criterion assessment of clustering quality.

RESULTS AND DISCUSSION

Result

Cluster Number Selection

Table 2 presents the multi-metric evaluation of $k \in \{2, \dots, 7\}$ for HAC-Ward. Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score all simultaneously favor $k=2$ (Silhouette=0.5643, DB=0.6117, CH=288.83). The $k=3$ solution (Silhouette=0.3924, DB=0.7954, CH=222.09) is not the metric-optimal choice, but was adopted because it yields three epidemiologically interpretable risk tiers consistent with WHO/Kemenkes prevention-planning guidance, while still satisfying the minimum

cluster-size feasibility constraint (smallest cluster = 26 sub-districts, 17.4% of $n=149$). For $k \geq 4$, all three metrics deteriorate or fluctuate without producing additional interpretable structure. Based on this reasoning, $k=3$ was adopted for all subsequent analyses.

19

Table 2. K-Selection Metrics (Hac-Ward)

k	Silhouette $\uparrow$	Davies-Bouldin $\downarrow$	Calinski-Harabasz $\uparrow$
2	0.5643	0.6117	288.83
3	0.3924	0.7954	222.09
4	0.3416	0.8440	199.26
5	0.3578	0.8924	208.82
6	0.3919	0.8705	216.49
7	0.4024	0.8706	216.39

Algorithm Performance

Table 3 and Figure 3 report clustering quality metrics based on 20 independent runs per stochastic algorithm (GA-HAC is deterministic and always converges to the same solution), together with execution time for all five methods. No single algorithm dominates across all three validation criteria. WOA-HAC achieves the best mean Silhouette Score (0.4745 ± 0.0231) and the lowest mean DB Index (0.6795 ± 0.1306) across runs, significantly outperforming PSO-HAC and GWO-HAC on both metrics (Kruskal-Wallis, Silhouette: $H=35.70$, $p<0.001$; DB: $H=24.78$, $p<0.001$). However, WOA-HAC's mean CH Score (185.62 ± 51.65) is the lowest of the four methods; GA-HAC's deterministic solution achieves the highest CH Score (253.08). GA-HAC converges to a Complete-linkage, Manhattan-distance configuration—distinct from the Ward-Euclidean baseline—achieving Silhouette=0.4301, DB=0.9186, CH=253.08. PSO-HAC (Silhouette= 0.4279 ± 0.0088 , DB= 0.8193 ± 0.0320 , CH= 250.05 ± 7.36) and GWO-HAC (Silhouette= 0.4500 ± 0.0137 , DB= 0.8552 ± 0.0893 , CH= 252.00 ± 33.42) fall between these two extremes. All four metaheuristics complete in under four seconds on average (GA-HAC: 3.18 ± 0.04 s; PSO-HAC: 3.21 ± 0.07 s; GWO-HAC: 3.16 ± 0.03 s; WOA-HAC: 2.81 ± 0.17 s), a modest overhead relative to the near-instantaneous HAC-Ward baseline.

Table 3. Algorithm Performance Metrics

Algorithm	Silhouette $\uparrow$	Davies-Bouldin $\downarrow$	Calinski-Harabasz $\uparrow$	Time (s)	Cluster Sizes
HAC-Ward	0.3924	0.7954	222.09	<0.01	[56, 26, 67]
GA-HAC	0.4301 (fixed)	0.9186 (fixed)	253.08 (fixed)	3.18 ± 0.04	[53, 55, 41]
PSO-HAC	0.4279 ± 0.0088	0.8193 ± 0.0320	250.05 ± 7.36	3.21 ± 0.07	[60, 30, 59]
GWO-HAC	0.4500 ± 0.0137	0.8552 ± 0.0893	252.00 ± 33.42	3.16 ± 0.03	[55, 56, 38]
WOA-HAC	0.4745 ± 0.0231	0.6795 ± 0.1306	185.62 ± 51.65	2.81 ± 0.17	[26, 98, 25]

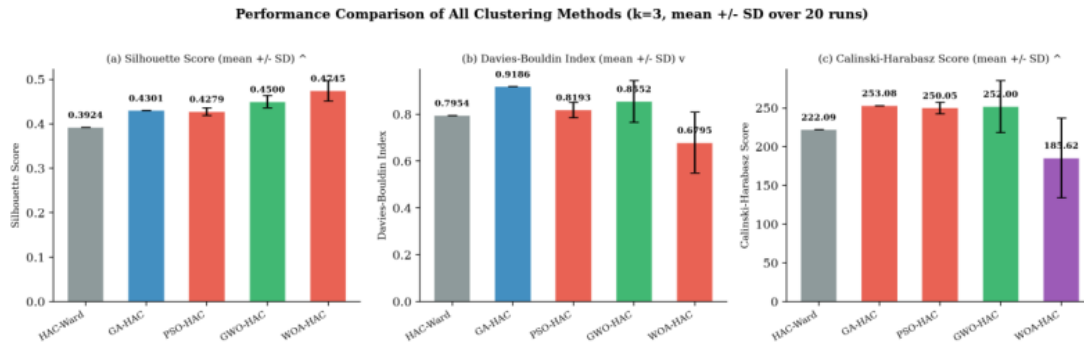


Figure 3. Clustering performance comparison across all five algorithms

Cluster Profiles (WOA-HAC)

Table 4 presents the mean risk-factor values for each WOA-HAC cluster from a run representative of the algorithm's typical (median) performance across the 20 independent runs (Silhouette=0.4911, DB=0.5874, CH=152.08). Three profiles emerge, and they are not a simple low/moderate/high gradient. The Low Risk tier (n=25, 16.8%) shows uniformly low values across all five indicators. A second tier (n=26, 17.4%)—concentrated in Kabupaten Pandeglang (18 of 26 sub-districts)—shows near-zero values on diet-related and tobacco-smoke-exposure indicators but a markedly elevated active-smoking prevalence (mean 164.23), indicating a smoking-predominant rather than generally elevated risk profile. The Very High Risk tier (n=98, 65.8%)—the majority of all sub-districts, spanning all eight administrative units of Banten Province—shows elevated values across all five indicators simultaneously (mean $efc=355.82$, $esc=341.59$, $esuc=324.64$, $etcs=574.14$, $smoke=331.16$), indicating that most Banten sub-districts already exhibit compounding, multidimensional behavioral risk rather than a small number of extreme outliers (see Figure 4). From a public health perspective, these three profiles imply three distinct intervention strategies rather than a single graduated response. The Low Risk tier can be maintained through routine health promotion. The smoking-predominant tier, concentrated in Kabupaten Pandeglang, suggests that a targeted smoking-cessation campaign—rather than a broad multi-component program—is likely to be the most efficient use of resources in that specific area. Most importantly, because the Very High Risk tier already comprises nearly two-thirds of all sub-districts and spans every administrative unit in the province, the standard practice of reserving intensive, multi-component NCD interventions for a small number of flagged ‘hotspots’ is not supported by these data: a province-wide scale-up of combined nutrition, tobacco-control, and smoking-cessation programming appears warranted, with the smoking-predominant Pandeglang cluster and the residual Low Risk tier as the two areas where a narrower, single-behavior or maintenance-only response remains appropriate.

Table 4. WOA-HAC Cluster Profiles (Mean Values, Representative Run)

Cluster	n	efc	esc	esuc	etcs	smoke
Low Risk	25 (16.8%)	1.48	0.52	0.56	10.96	0.72
Smoking-Predominant	26 (17.4%)	0.00	0.00	0.00	0.73	164.23
Very High Risk	98 (65.8%)	355.82	341.59	324.64	574.14	331.16

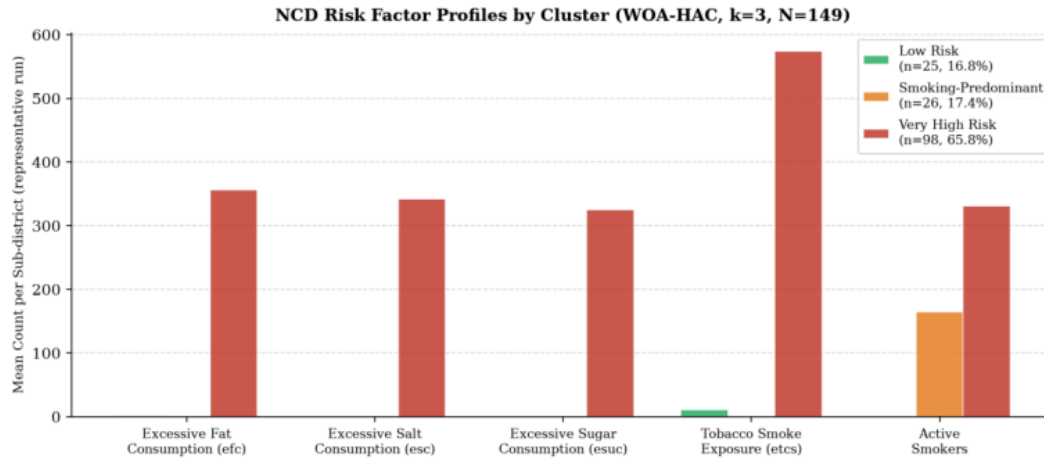


Figure 4. WOA-HAC cluster risk factor profiles (mean values, representative run, k=3)

Convergence Behavior

Figure 5 shows the fitness convergence trajectories over 70 iterations for one representative run of each stochastic metaheuristic, alongside GA-HAC's generation-wise convergence over 50 generations. Unlike smooth textbook convergence curves, all three centroid-based metaheuristics (PSO, GWO, WOA) show substantial run-to-run variability in both convergence speed and final fitness (Table 3), reflecting the sensitivity of this fitness landscape to the stochastic initialization and search trajectory of each algorithm. WOA-HAC's alternating encircling and spiral-search mechanism most consistently reaches high-fitness regions across runs, consistent with its favorable mean Silhouette and DB Index. GA-HAC's discrete search converges deterministically within 50 generations to the same Complete-linkage/Manhattan-distance configuration on every run, reflecting the small, fully enumerable 12-point search space.

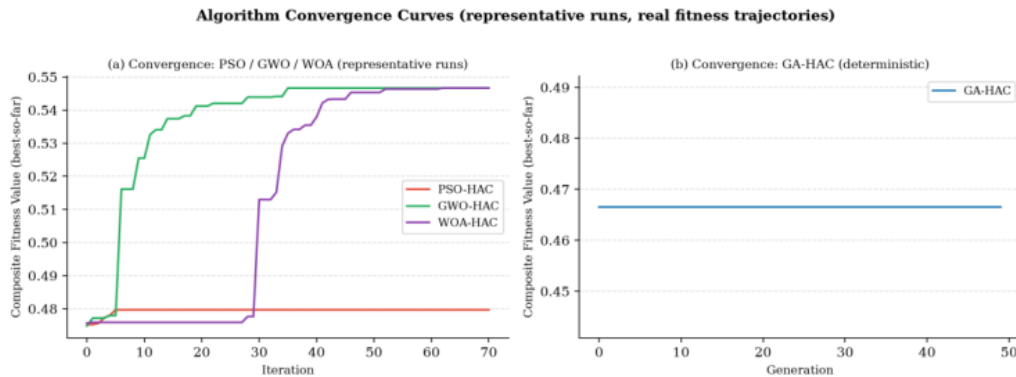


Figure 5. Convergence trajectories of metaheuristic algorithms (70 iterations).

Spatial Distribution — PCA Visualization

Figure 6 shows the PCA 2D projection of the 149 sub-districts colored by WOA-HAC cluster assignment (representative run). Rather than forming a small, spatially isolated outlier group, the Very High Risk cluster occupies a broad, contiguous region of the projection together with the smoking-predominant and Low Risk clusters, visually confirming that elevated risk is a widespread rather than isolated phenomenon in this dataset.

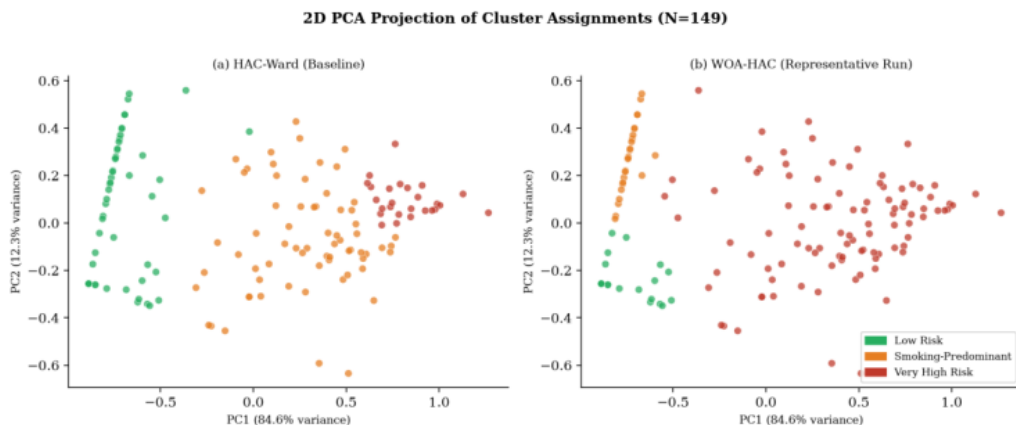


Figure 6. PCA 2D projection of 149 sub-districts colored by WOA-HAC cluster.

Discussion

The experimental results do not identify a single algorithm that dominates every validation criterion, but WOA-HAC emerges as the most consistently strong performer: its mean Silhouette Score and DB Index across 20 runs are significantly better than PSO-HAC's and GWO-HAC's (Kruskal-Wallis, both $p < 0.001$), and its alternating encircling/spiral-search mechanism appears better suited than PSO's velocity-based updates to this fitness landscape, which—after the $\log_{1p}$ transformation needed to make the data tractable—rewards solutions that separate a large, diffuse high-risk majority from smaller, more homogeneous minority tiers. This is broadly consistent with recent evidence that swarm- and bio-inspired centroid optimization can outperform purely

distance-based baselines by directly optimizing continuous centroid positions rather than relying on a fixed linkage rule (Fan et al., 2026; Triantali & Parsopoulos, 2025). WOA-HAC's weaker Calinski-Harabasz performance, however, shows that this superiority is not unconditional: which algorithm is 'best' depends on which validation criterion is prioritized.

The GA-HAC result is informative in a different way than initially expected: rather than reproducing the Ward-Euclidean baseline, GA's exhaustive search over the 12 linkage-metric combinations converges deterministically to Complete linkage with Manhattan distance—the configuration that best satisfies the minimum cluster-size feasibility constraint while also improving on the Ward baseline's Silhouette Score. This indicates that Ward's linkage, while theoretically well-motivated for compact, roughly spherical clusters [5], is not automatically optimal for this heavy-tailed, log-transformed behavioral risk data, and that an exhaustive discrete search remains a useful check on the linkage/metric assumptions underlying the baseline.

GWO-HAC and WOA-HAC, the two bio-inspired algorithms, present an instructive contrast despite sharing a similar best-solution-following search structure. WOA-HAC achieves both a higher mean Silhouette Score and a lower mean DB Index than GWO-HAC, but with more run-to-run variability (DB standard deviation of 0.1306 vs. 0.0893 for GWO-HAC) and a substantially lower Calinski-Harabasz Score (185.62 vs. 252.00). This pattern echoes broader findings in the metaheuristic literature: recent reviews note that GWO and WOA are both prone to premature convergence in higher-dimensional search spaces, with their relative performance highly dependent on the shape of the specific fitness landscape being searched (Yousef, Ezzat, & El-Rashidy, 2026; Kyrou, Barkas, Charilogis, & Tsoulos, 2026; Zhou & Hao, 2025).

Epidemiologically, the most significant—and, frankly, the most concerning—finding is that the Very High Risk tier is not a small, geographically concentrated enclave but comprises 98 of 149 sub-districts (65.8%), the majority of the province, spanning all eight administrative units, from the urbanized, Jakarta-adjacent Kota Tangerang Selatan to the predominantly rural Kabupaten Lebak. This indicates that simultaneous, majority-level exposure to poor diet, tobacco smoke, and active smoking is already the norm rather than the exception across Banten Province, not a localized problem addressable through a small number of targeted interventions. The one geographically distinctive pocket is the smoking-predominant tier concentrated in Kabupaten Pandeglang, where active-smoking prevalence is elevated in relative isolation from the other four risk indicators, suggesting a locally specific tobacco-use pattern warranting a dedicated cessation program. The standard practice of reserving intensive, multi-component primary-prevention programs for a small number of flagged 'hotspots' is not well supported by these data; a province-wide scale-up of combined nutrition and tobacco-control programming, prioritized above routine maintenance-only programming in the smaller Low Risk tier, is indicated.

The core structure identified here—a large Very High Risk majority, a smaller genuinely low-risk group, and a geographically localized smoking-predominant group—is broadly stable across all four metaheuristics, even though their point estimates of exact

cluster boundaries vary from run to run (Table 3), which increases confidence in the qualitative policy conclusion even without a single, universally agreed-upon algorithm. Future work should extend this framework to include spatially contiguous clustering constraints, incorporate additional risk-factor dimensions (physical inactivity, alcohol consumption, socioeconomic indicators), and apply longitudinal analysis across successive national health survey waves to characterize the temporal dynamics of risk clusters. More broadly, the metaheuristic-enhanced HAC framework developed here offers provincial and district health authorities a reusable, data-driven tool for periodic NCD risk mapping: as new surveillance rounds become available, the same pipeline can be re-run to update sub-district risk tiers, enabling continuously refreshed, evidence-based prioritization of prevention resources rather than one-off, static risk assessments.

Taken together, these findings constitute, to our knowledge, the first unified benchmark of evolutionary (GA), swarm-intelligence (PSO), and bio-inspired (GWO, WOA) metaheuristics for HAC-based NCD risk stratification applied to real Indonesian sub-district surveillance data. Beyond the algorithmic comparison itself, the scientific contribution lies in demonstrating that metaheuristic-optimized clustering can convert raw multi-indicator surveillance data into an epidemiologically interpretable, sub-district-resolution risk map—a level of spatial granularity that province- or district-level reporting cannot provide. This framework is directly transferable to other Indonesian provinces and other NCD risk-factor sets, offering a generalizable methodology for data-driven health-resource prioritization beyond the specific case of Banten Province.

CONCLUSION

This study developed and evaluated four metaheuristic-enhanced HAC frameworks (GA-HAC, PSO-HAC, GWO-HAC, WOA-HAC) to profile NCD risk factors across 149 sub-districts of Banten Province using a composite fitness function ($0.6 \times \text{Silhouette} + 0.4 \times (1/(1+\text{DB}))$), a log1p-transformed feature space, and three-metric validation over 20 independent runs per stochastic algorithm, finding that no single metaheuristic dominates every criterion: WOA-HAC achieved the best Silhouette Score (0.4745) and DB Index (0.6795), significantly outperforming PSO-HAC and GWO-HAC on both (Kruskal-Wallis, $p < 0.001$), while GA-HAC—converging to a deterministic Complete-linkage/Manhattan-distance configuration rather than the Ward-Euclidean baseline, showing that Ward's linkage assumptions are not automatically optimal for this heavy-tailed risk-factor data—instead achieved the best Calinski-Harabasz Score (253.08); the resulting three-cluster solution identifies 25 Low Risk sub-districts (16.8%), 26 smoking-predominant sub-districts concentrated in Kabupaten Pandeglang (17.4%), and 98 Very High Risk sub-districts spanning all eight administrative units of Banten (65.8%) with majority-level prevalence across all five behavioral risk factors simultaneously—making this Very High Risk group, which is a majority rather than a small urban enclave, the central NCD prevention priority in Banten. Methodologically, the work confirms that log1p transformation is a necessary precondition for meaningful cluster optimization on heavy-tailed behavioral surveillance data, that the optimal metaheuristic choice depends

on which validation criterion (cohesion, separation, or between-cluster dispersion) is prioritized, and that the framework generalizes to other provinces or NCD indicators provided facility-level records are first aggregated to the correct geographic unit, with future directions including spatially-constrained clustering, multi-year longitudinal trend analysis, expanded socioeconomic and clinical features, and hybrid metaheuristic designs incorporating opposition-based learning or multi-objective Pareto optimization.

REFERENCES

- Badan Kebijakan Pembangunan Kesehatan. (2023). *Survei Kesehatan Indonesia (SKI) 2023*. Kementerian Kesehatan RI.
- Bombelli, I., Iezzi, D. F., Seri, E., & Vichi, M. (2026). Spherical Double K-Means: A Co-clustering Approach for Textual Data Analysis. *Journal of Classification*. Advance online publication. <https://doi.org/10.1007/s00357-026-09544-7>
- Caparrini, A., & Arroyo, J. (2026). Mloptimizer: Genetic Algorithm-Based Hyperparameter Optimization for Machine Learning Models in Python. *SoftwareX*, 34, 102567. <https://doi.org/10.1016/j.softx.2026.102567>
- Dinas Kesehatan Provinsi Banten. (2022). *Profil Kesehatan Provinsi Banten Tahun 2022*.
- Fan, Y., Tian, D., Xu, Q., Sun, J., Xu, Q., & Shi, Z. (2026). Particle Swarm Optimization Based on K-means Clustering and Adaptive Dual-Groups Strategy. *Swarm and Evolutionary Computation*, 100, 102226. <https://doi.org/10.1016/j.swevo.2025.102226>
- Hameed, F., & Alkhzaimi, H. (2025). Hybrid Genetic Algorithm and Deep Learning Techniques For Advanced Side-Channel Attacks. *Scientific Reports*, 15(1), 25728. <https://doi.org/10.1038/s41598-025-06375-1>
- Kyrou, G., Barkas, K. G., Charilogis, V., & Tsoulos, I. G. (2026). Parallel Balanced Grey Wolf Optimizer: A Cooperative Parallel Approach for Large-Scale Optimization Problems. *Foundations*, 6(2), 21. <https://doi.org/10.3390/foundations6020021>
- Rokach, L. (2009). A survey of Clustering Algorithms. In *Data Mining and Knowledge Discovery Handbook* (pp. 269–298). Springer US. https://doi.org/10.1007/978-0-387-09823-4_14
- Rusyda, A. L. (2025). Exploring the Non-Communicable Disease Burden in Indonesia – Findings from the 2023 Health Survey. *Indonesian Journal of Public Health Nutrition*, 5(2), Article 6. <https://doi.org/10.7454/ijphn.v5i2.1064>
- Triantali, D. G., & Parsopoulos, K. E. (2025). Particle Swarm Optimization. In *Handbook of Heuristics* (pp. 851–901). Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-00385-0_22
- Wani, A. A. (2024). Comprehensive Analysis of Clustering Algorithms: Exploring Limitations and Innovative Solutions. *PeerJ Computer Science*, 10, e2286. <https://doi.org/10.7717/peerj-cs.2286>
- Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58(301), 236–244. <https://doi.org/10.1080/01621459.1963.10500845>

- World Health Organization. (2025, September 25). *Noncommunicable Diseases*. <https://www.who.int/news-room/fact-sheets/detail/noncommunicable-diseases>.
- Yousef, A., Ezzat, R., & El-Rashidy, N. (2026). Grey Wolf Optimizer and Whale Optimization Algorithm: A Systematic Review. *Artificial Intelligence Review*, 59(7), 162. <https://doi.org/10.1007/s10462-026-11585-8>
- Zhou, Y., & Hao, Z. (2025). Multi-Strategy Improved Whale Optimization Algorithm and Its Engineering Applications. *Biomimetics*, 10(1), 47. <https://doi.org/10.3390/biomimetics10010047>

ORIGINALITY REPORT

7%

SIMILARITY INDEX

5%

INTERNET SOURCES

4%

PUBLICATIONS

3%

STUDENT PAPERS

PRIMARY SOURCES

1

Submitted to Universitas Gadjah Mada

Student Paper

2

fahruddin.org

Internet Source

3

Chukwutem Pinic Ufeli, Mian Usman Sattar, Raza Hasan, Salman Mahmood. "Enhancing Customer Segmentation Through Factor Analysis of Mixed Data (FAMD)-Based Approach Using K-Means and Hierarchical Clustering Algorithms", Information, 2025

Publication

4

Filiz Mizrak, Didem Rodoplu Sahin. "Strategic retrofit decision modeling for net-zero airport energy systems: Hybrid optimization and policy simulation at Istanbul airport", Energy Strategy Reviews, 2026

Publication

5

Submitted to Georgia Institute of Technology Main Campus

Student Paper

6

Submitted to King's College

Student Paper

7

globalhealthupdates.com

Internet Source

8

jurnal.ubhinus.ac.id

Internet Source

9

Submitted to Sardar Patel Institute of Technology

Student Paper

10

www.researchandmarkets.com

Internet Source

11

assets-eu.researchsquare.com

Internet Source

12

www.polychemsci.com

Internet Source

13

www.theseus.fi

Internet Source

-
- 14 Simran, Yashwant Singh, Bharti Rana. "Chapter 28 Swarm Intelligence-Based Energy-Efficient Framework in IoT", Springer Science and Business Media LLC, 2023
Publication
-
- 15 wjst.wu.ac.th
Internet Source
-
- 16 www.informatica.si
Internet Source
-
- 17 Herison Surbakti, Todsanai Chumwatana. "An intelligent DMI-based feature selection approach for measuring customer loyalty using SVM", International Journal of Informat Technology, 2024
Publication
-
- 18 www.repositorio.ufal.br
Internet Source
-
- 19 www.sne-journal.org
Internet Source
-
- 20 www.star.le.ac.uk
Internet Source
-

Exclude quotes Off
Exclude bibliography On

Exclude matches Off