JURNAL MATEMATIKA, STATISTIKA DAN KOMPUTASI e-ISSN: 2614-8811
Published by Departement of Mathematics, Hasanuddin University, Indonesia
https://journal.unhas.ac.id/index.php/jmsk/index
Vol. 22, No. 3, May 2026, pp. 761-780
DOI: 10.20956/j.v22i3.50288

Optimizing Cluster Centroids using Hybrid Firefly-Genetic Algorithm for Village Development Clustering

Annisa Rahma[1]*, Rani Nooraeni[2], Raditya Hizra Maharani[3]

[1]*Directorate of Services Statistics, Badan Pusat Statistik RI, Jakarta, Indonesia*
[2,3]*Departement of Statistical Computing, Politeknik Statistika STIS, Jakarta, Indonesia*
***Email**: [1]annisarahma@bps.go.id, [2]raninoor@stis.ac.id, [3]222212824@stis.ac.id*
**Corresponding author*
Received: 12 March 2026, revised: 6 May 2026, accepted: 7 May 2026

Abstract

Clustering in the Building Village Index (BVI) offers an alternative approach to identifying village groupings based on numerical and categorical characteristics similarities. K-Prototype (KP) is a popular clustering algorithm for handling mixed numerical and categorical data. However, it tends to converge to a local optimum due to random centroid initialization. Metaheuristic approaches have been widely applied to improve centroid initialization, yet simple metaheuristics often yield suboptimal solutions due to limited exploration or exploitation capabilities. This study proposes the hybridization that combines the exploitation strength of the Firefly Algorithm (FA) with the exploration capability of the Genetic Algorithm (GA) to optimize centroid initialization of KP. The results show that FA-GA hybridization enables the centroid initialization process to be faster with optimal fitness compared to either a single FA or a single GA. FGAKP is the best clustering algorithm in this study because it produces the smallest Total Cost (TC) and the largest Cluster Validity (CV) index with the most efficient centroid initialization time across all training data. The implementation of FGAKP in village grouping based on BVI indicators in North Kalimantan Province in 2024 grouped 484 villages into 5 categories based on their village development potential, achieving an 11.69% reduction in TC and a 16.75% improvement in CV compared to the standard KP without optimization.

Keywords: *village grouping, k-prototype, firefly algorithm, genetic algorithm, hybridization*

1. INTRODUCTION AND PRELIMINARIES

The rapid development of data and technology has made business organizations face increasingly complex decision-making challenges [23]. As a result, the availability of diverse data in large quantities necessitates effective processing to produce useful information for informed decision-making [32]. In response to these challenges, a measure was developed to strengthen development at the village level by capturing the growth of village independence through the Building Village Index (BVI) [8], [11]. In calculating BVI, each variable is scored, with indicators determined at the national level and uniformly applied across all villages [13]. However, such

standardized weightings may not be relevant to the diverse conditions in individual villages, thereby limiting the scoring method's ability to capture specific characteristics and underlying heterogeneity patterns. Consequently, this may reduce the accuracy and reliability of the resulting decision-making process. To address this need, the implementation of clustering in BVI serves as an alternative for identifying grouping patterns of villages based on region-specific characteristics such as development level, accessibility, or socio-economic conditions, without the need for assigning scores or labels. Through clustering, it is possible to analyze pattems, relationships, and uncover hidden information in the data [21].

In 2024, North Kalimantan was recorded as the province with the lowest provincial-level BVI and the only province on the island of Kalimantan that is below the national average [12]. North Kalimantan has a strategic position as a logistical, ecological, and economic buffer area for the IKN. This role requires regional readiness, particularly at the village level, to support sustainable regional development. Therefore, a comprehensive study is needed regarding the condition of village progress and independence in North Kalimantan, particularly through village grouping based on the Building Village Index (BVI) indicator, as an effort to map regional development in the Kalimantan region.

The primary source of BVI is the Potensi Desa (Podes) data [11]. The Podes dataset comprises hundreds of attributes, including both numerical and categorical variables, with villages serving as the unit of analysis at the lowest administrative level. Based on these characteristics, Podes' data is a complex dataset, so it requires an efficient approach to find optimal solutions in analyzing the characteristics of a region. K-Prototype is a partitioning-based clustering method that can handle mixed numerical and categorical data [27]. Nonetheless, the K-Prototype algorithm often converges on local optima rather than on global optima [19]. In addition, K-Prototypes are sensitive to the determination of cluster centroid initialization and tend to undergo premature convergence, resulting in local optimum solutions. As a consequence, clustering outcomes can vary significantly depending on different random initializations of cluster centers [15]. This highlights a fundamental challenge in K-Prototypes clustering, namely the difficulty in obtaining stable and robust centroid initialization.

Metaheuristic algorithms can be applied to enhance the efficiency and accuracy of machine learning across various tasks, including clustering, by optimizing the initialization of centroids [20], [25]. For example, the Firefly Algorithm (FA) is a metaheuristic algorithm that can perform automatic subdivisions and handle multimodality [29]. It enables effective centroid initialization by simultaneously exploring multiple potential cluster centers. Empirical evidence from Senthilnath et al. [24] shows that FA is more effective than 11 other methods for clustering in most cases. The FA algorithm is mighty in local searches (exploitation), but it can become stuck in some local optima, which limits its ability to perform a global search effectively [5]. In addition, FA employs a full attraction mechanism in which each firefly interacts pairwise with all other fireflies, resulting in a high computational cost with a time complexity of $O(N^2)$ [33]. In contrast, the Genetic Algorithm (GA) is a metaheuristic algorithm that possesses excellent global search (exploration) capabilities, enabling it to avoid the trap of local optima [26]. GA also uses simple genetic operators—selection, crossover, and mutation—each with linear-time complexity $O(N)$ relative to the population size [7]. However, GA tends to experience premature convergence due to low exploitation of diversity in populations. Existing studies mainly focus on standalone metaheuristic approaches which often struggle to balance exploration and exploitation [1], particularly in optimizing centroid initialization for K-Prototype clustering with mixed numerical and categorical data. Therefore, hybridizing FA and GA offers a promising solution for optimization problems of a single or simple metaheuristic [2], [4]. By integrating both algorithms' strengths, the researcher combines FA and GA to develop a hybrid algorithm that delivers a more effective and efficient solution of centroid initialization process.

The metaheuristic approach can be used as an alternative to improve the K-Prototypes algorithm [27]. According to Kaur et al. [10], the metaheuristic approach can be used to iterate solution candidates that improve the fitness value and quality of clusters from the K-Prototypes algorithm. Furthermore, the metaheuristic algorithm is capable of handling non-convex clusters through complex search space exploration and nonlinear determination of cluster boundaries. This study proposes a hybrid FA and GA approach to optimize centroid initialization in K-Prototype clustering, aiming to achieve a better balance between exploration-exploitation trade-off and improve overall clustering performance. The best-performing method is then applied to identify the pattern of grouping conditions of village development and independence based on the Building Village Index (BVI) indicator.

2. MATERIALS & METHODS

2.1 Data and Data Sources

The benchmark datasets of this study refer to research by Nooraeni and Nurfalah [16] namely Zoo, Acute Inflammations, and Credit Approval. These datasets were selected because they consist of mixed-type attributes (categorical and numerical) and possess a sufficient number of observations. The Zoo dataset comprises 1 numeric and 15 categorical attributes with 101 observations; the Acute Inflammations dataset consists 1 numeric and 5 categorical attributes with 120 observations; and the Credit Approval dataset features 6 numeric and 9 categorical attributes with 690 observations. To extend the validation, the Heart Disease dataset from UCI Machine Learning Repository that consists of 6 numeric and 7 categorical attributes with 303 observations was also included [9].

Following the validation on these benchmark datasets, the proposed method was implemented on official real-world data, specifically the Village Development Data (Podes). The data was based on the BVI indicator, which were adjusted to match the available data in the Podes dataset. The Podes dataset consists of 17 numeric and 23 categorical attributes with 484 observations. A summary of the datasets for this research provided in Table 2.1.

Table 2.1. Summary of Benchmark and Village Developement Datasets

Dataset	Numeric Attributes	Categorical Attributes	Observations
Zoo	1	15	101
Acute Inflammations	1	5	120
Credit Approval	6	9	690
Heart Disease	6	7	303
Village Development Data	17	23	484

2.2 Clustering Mixed-type Data using Metaheuristic Algorithms

This study employs 5 types of clustering algorithms: K-Prototype (KP), Firefly Algorithm KP (FAKP), Genetic Algorithm KP (GAKP), Hybrid Firefly-Genetic Algorithm KP (FGAKP), and Hybrid Genetic-Firefly Algorithm KP (GFAKP). These 5 algorithms were applied to the trial data to identify the best clustering algorithm, which was then implemented in the village grouping in North Kalimantan Province based on BVI indicators using Podes 2024 data. In the Firefly Algorithm (FA), the parameters that need to be tuned include the number of fireflies (n), attractiveness coefficient (β_0), light absorption coefficient (γ), cooling factor (δ), and maximum iterations ($iter_{max}$). The values are adopted from previous research [30], including n = 50, β_0 = 1, γ = 1, δ = 0.95, and $iter_{max}$ = 30. Meanwhile, in the Genetic Algorithm (GA), the parameters

that need to be tuned include the number of chromosomes (N), crossover rate (P_c), mutation rate (P_m), and maximum iterations (Gen_{max}). The values are adopted from previous research (Nooraeni, 2016), including N = 50, P_c = 0.7, P_m = 0.01, and Gen_{max} = 30. To ensure a fair comparison and equal computational cost across algorithms, the hybrid algorithm uses the same tuning parameters as the single algorithms. However, the maximum number of iterations is split equally between the two stages, allocating 15 iterations to the FA stage ($iter_{max}$ = 15) and 15 generations to the GA stage (Gen_{max} = 15), following the approach proposed in Elkhechafi et al. [3] and Nand and Sharma [14] research.

K-Prototype (KP)

There are four main stages in the KP algorithm where the number of clusters to be formed has been determined beforehand [28]:

1. Determine number of clusters (L) with cluster initialization $(z_1, z_2, \dots, z_L)$ from the n point $\{x_1, x_2, \dots, x_n\}$
2. Calculate the distance of all data points/objects to the initials of the initial cluster by the following equation (2.1) and allocate all data points into clusters with the closest distance. The total distance in equation (2.1) corresponds to the total cost in the K Prototype algorithm.

$$d(X_i, Z_l) = \left(\sum_{j=1}^{m_r} \left(x_{ij}^r - z_{lj}^r\right)^2 + \lambda \sum_{j=1}^{m_c} \delta(x_{ij}^c, z_{lj}^c) \right)^{\frac{1}{2}} \tag{2.1}$$

3. Calculate the new cluster centroids for each cluster. After that, relocate all data points in the dataset to the new centroid/prototype;
4. Check that the termination condition (no centroid change) is met. If not, return to stages 2 and 3 and repeat the process until the maximum iteration is reached.

To optimize the performance of the K-Prototypes algorithm, this study proposes four experimental scenarios focused on initial centroid optimization. The first two scenarios utilize the Firefly Algorithm (FA) and Genetic Algorithm (GA) as standalone optimizers. Building on these, the third and fourth scenarios employ hybrid strategies: the former integrates FA followed by GA (FA-GA) while the letter reverses this sequence to employ a GA-FA approach.

Scenario 1: Initializing Centroids with hybrid Firefly Algorithm K-Prototype (FAKP)

The steps in the FA algorithm are explained below [18]:

1. Initialize the parameters n, β_0, γ, δ, and $iter_{max}$;
2. Generate N random initial population of fireflies;
3. Calculate the light intensity of firefly (I) using equation (2.2) as follows:

$$I = f(x) \tag{2.2}$$

$f(x)$ is a fitness value function $(\frac{1}{(1+h)})$ where h is the total cost function of the KP which refers to equation (2.1);

4. Count the movements of firefly i, attracted to brighter firefly j, using equations (2.3) and (2.4) [29]:

$$x_i^{t+1} = x_i^t + \beta^0 e^{-\gamma r_{ab}^2}\left(x_j^t - x_i^t\right) + \alpha\left(rand - \frac{1}{2}\right) \tag{2.3}$$

$$\alpha = \alpha_0 \delta^t \quad (2.4)$$

For the categorical data, the characteristics of dimmer fireflies will follow the characteristics of brighter fireflies;

5. Check that the termination condition (no changes 10 times in a row) has been met. If not, return to stage 3 and repeat the process until the maximum iteration is reached;
6. Fireflies are produced with an optimal solution, which will then be the centroid initialization in KP.

Scenario 2: Initializing Centroids with K-Prototype Genetic Algorithm (GAKP)

The steps in the GA algorithm are explained below [15], [28]:

1. Initialize the parameters N, P_c, P_m, and Gen_{max};
2. Generate N random initial population of chromosomes;
3. Evaluating the fitness value of each individual (chromosomes);
4. Select the parent individual using the Roullete Wheel technique with equation (2.5):

$$P_x = \frac{f_x}{f_{total}} \quad x = 1,2,\dots,N \quad (2.5)$$

5. Apply crossover operations with Single-Point crossover and mutations with Swap Mutations on selected broods to produce offspring (new individuals). Then, select these parents and offspring as the final individuals.
6. Check the condition of termination (no change in fitness 10 times in a row or maximum iteration). If not met, return to step (3). If met continue to (7);
7. An individual is generated with an optimal solution, which will then be the centroid initialization in KP.

Scenario 3: Initializing the Centroid with the Firefly-Genetic Algorithm K-Prototype (FGAKP)

FGAKP employs a hybrid metaheuristic algorithm combining FA and GA. First, the FA algorithm is run, followed by GA. The FGAKP stages are described as follows [Appendix 1.]:

1. Initialize the parameters n, N, β_0, γ, δ, P_c, P_m, $iter_{max}$, and Gen_{max} ;
2. Generate an initial population of 50 fireflies at random;
3. Calculate I using equation (2.2);
4. Calculate the movement of the fireflies with equation (2.3);
5. Check the termination condition. If not, return to step 3 and repeat until the maximum iteration is reached.
6. The best solution from the FA algorithm that will be the initial population in GA;
7. Evaluate the fitness value of each chromosomes using the fitness function / objective function of KP in equation (2.1);
8. Apply the selection process using the Roulette Wheel technique;
9. After obtaining the selected chromosome, do the Single-Point crossover and Swap Mutation process;
10. Check the termination condition. If it is not fulfilled, return to step (7). If it is met, proceed to step (11);
11. Generated individuals with optimal solutions in the population that would be the centroid initialization of the KP.

Scenario 4: Initializing the Centroid with the Genetic-Firefly Algorithm K-Prototype (GFAKP)

GFAKP employs a hybrid metaheuristic algorithm combining FA and GA. First, the GA algorithm is run, followed by FA. The GFAKP stages are described as follows [Appendix 2.]:

1. Initialize the parameters n, N, β_0, γ, δ, P_c, P_m, $iter_{max}$, and Gen_{max};
2. Randomly generate an initial population of 50 chromosomes;
3. Evaluate the fitness value of each chromosomes using the fitness function / objective function of KP in equation (2.1);
4. Apply the selection process using the Roulette Wheel technique;
5. After obtaining the selected chromosome, do the Single-Point crossover and Swap Mutation process;
6. Check the termination condition. If it is not fulfilled, return to step (3). If it is met, proceed to step (7);
7. The best solution is generated from the GA algorithm that will be the initial population in FA;
8. Calculate I using equation (2.2);
9. Calculate the movement of the fireflies with equation (2.3);
10. Check the termination condition. If not, return to step 8 and repeat until the maximum iteration is reached.
11. Generated individuals with optimal solutions in the population that would be the initialization of the KP.

2.3 Determine the Proposed Method

In the centroid initialization process, fitness value and computational time were used as evaluation metrics. For evaluating clustering results across all applied algorithms (KP, FAKP, GAKP, FGAKP, and GFAKP), Total Cost (TC) and Cluster Validity Index (CV) are the main evaluation metrics. The Total Cost (TC) is the objective function that minimizes the distance between each data point and its respective cluster centroid [15]. Meanwhile, the Cluster Validity Index (CV) is used as a validity measure to assess the quality of the clustering structure. The higher the CV Index, the better the clustering results [6]. In addition, the number of iterations and computational time are employed as supporting performance metrics.

Total Cost (TC)

The total cost function equation for numerical and categorical mixed data can be expressed by equations (2.6) and (2.7) as follows [15]:

$$cost_l = \sum_{l=1}^{L} u_{il} \sum_{j=1}^{m_r} \left(x_{ij}^r - z_{lj}^r\right)^2 + \sigma_l \sum_{l=1}^{L} u_{il} \sum_{j=l+1}^{m_c} \delta(x_{ij}^c, z_{lj}^c) \tag{2.6}$$

$$Cost_l = Cost_l^r + Cost_l^c \tag{2.7}$$

Cluster Validity Index (CV)

The CV Index equation can be written with equation (2.8) as follows [6]:

$$IndeksCV = \frac{CU}{(1+\sigma^2)} \tag{2.8}$$

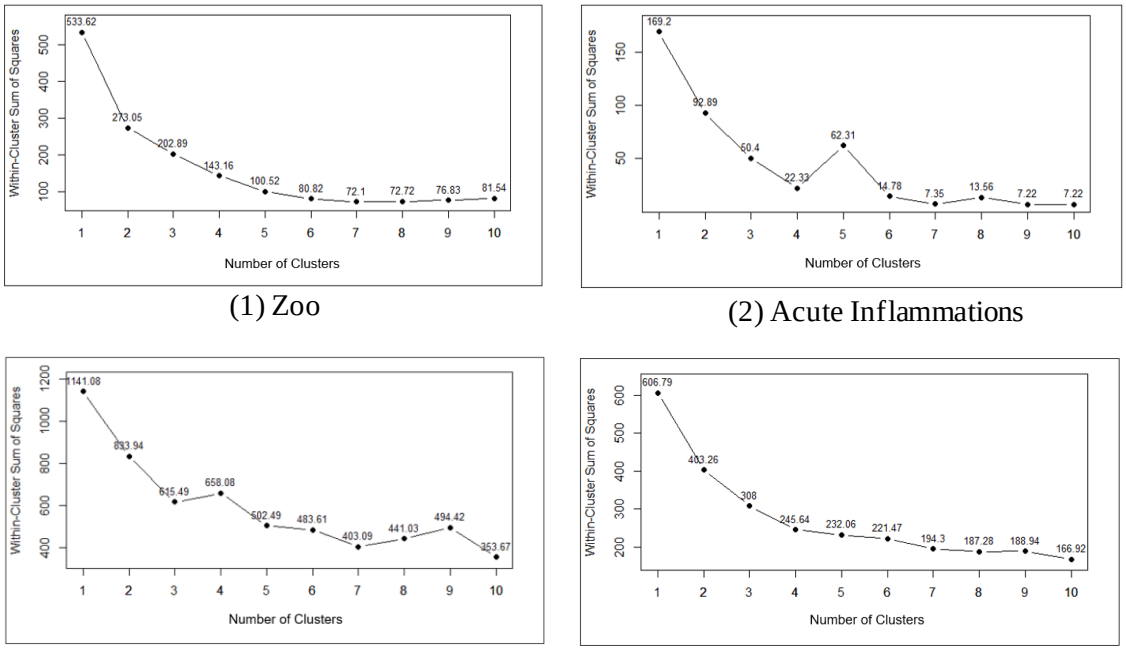
$$CU = \sum_{l} \left(\frac{|C_l|}{|D|} \sum_{j} \sum_{l} \left[P(C_l)^2 - P\left(A_j = V_{ij}\right)^2 \right] \right) \tag{2.9}$$

$$\sigma^2 = \sum_l \frac{1}{|C_l|} \sum_j \sum_l \left(V_{ij}^l - V_{j,avg}^l\right)^2 \tag{2.10}$$

3. CLUSTERING RESULT ON MIXED-TYPE DATASETS

3.1 Determination of the Optimal Number of Clusters

Based on the determination of the optimal number of clusters using the Elbow method in Figure 3.1 and considering the reference information on the number of classes, the number of clusters in the test data used in this study is: 6 clusters for Zoo, 4 clusters for Acute Inflammations, 3 clusters for Credit Approval, and 5 clusters for Heart Disease.



(1) Zoo (2) Acute Inflammations

(3) Credit Approval (4) Heart Disease

Figure 3.1. Determination of the Optimal Number of L Clusters on the Benchmark Datasets

For the case study, Village Development Data (Podes), it can be observed from Figure 4 that there was a significant decrease in WCSS until the fifth cluster, after which it began to slope. Therefore, in the case study data, this study utilizes as many clusters as possible.

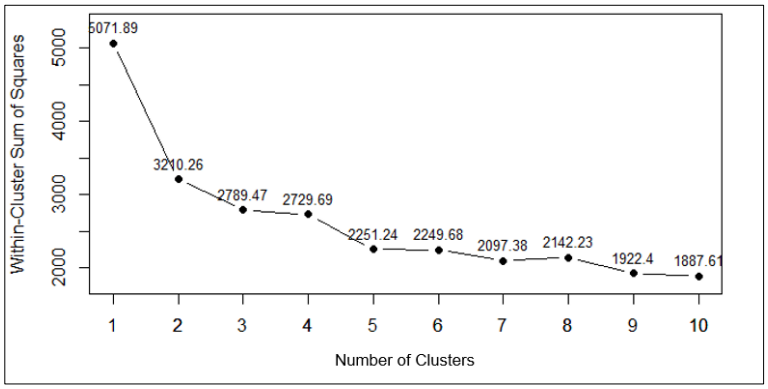
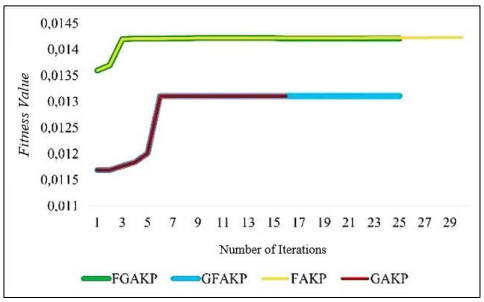


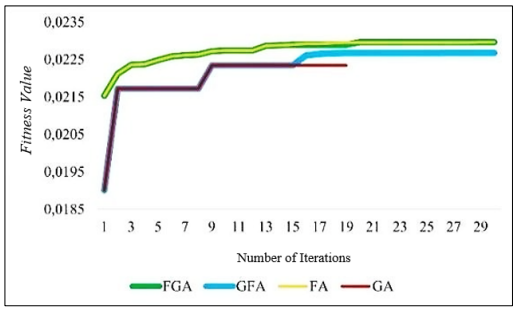
Figure 3.2. Determination of Optimal Cluster L on Case Study Data

3.2 Centroid Initialization on Benchmark Datasets with FAKP, GAKP, FGAKP, and GRAKP Algorithms

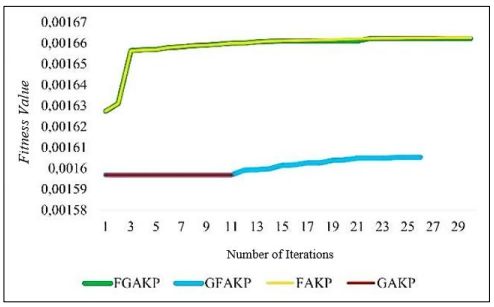
To determine the performance of each optimization algorithm in producing optimal centroid initialization in the test data, a comparison was made of the fitness value, number of iterations, and computational time of the optimization algorithm presented in Figure 3.3 and Table 3.1, as follows:



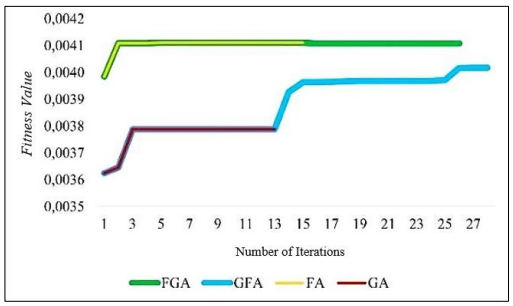
(1) Zoo



(2) Acute Inflammations



(3) Credit Approval



(4) Heart Disease

Figure 3.3. Fitness Value and Number of Iterations on Benchmark Data

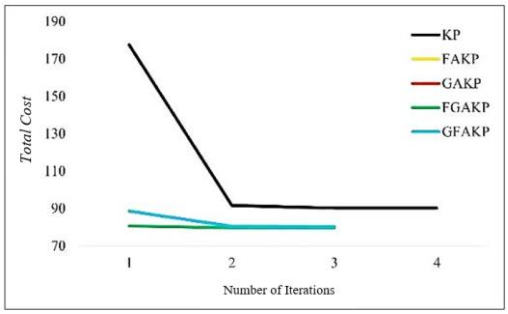
Table 3.1. Centroid Initialization Computation Time on Benchmark Datasets (in seconds)

Dataset	FAKP	GAKP	FGAKP	GFAKP
Zoo	22.712	0.233	6.978	5.981
Acute Inflammations	9.906	0.191	5.267	6.675
Credit Approval	95.470	0.724	31.449	32.004
Heart Disease	22.068	0.465	18.811	20.303

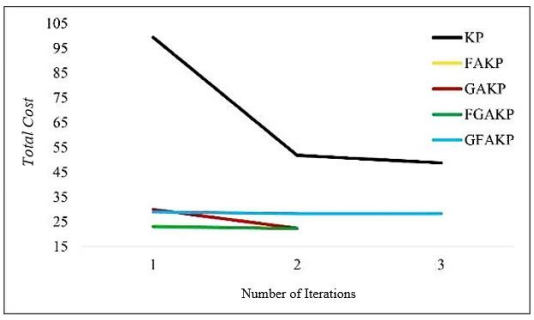
The FGAKP hybrid optimization algorithm in the Zoo dataset is 69.27% superior in terms of computational time to FAKP and 8.41% in terms of fitness value to GAKP. This is in line with Ezzeldin et al. [4] that states that hybrid algorithms of FA and GA produce models with faster convergence towards global optimum solutions and are able to achieve the best value fitness compared to their single algorithms. In the Acute Inflammations dataset, it can be observed that there is an increase in fitness value in GFAKP after the implementation of FA in a sequential manner. This is in line with the statement Nand and Sharma [14] that, in the combination of GA and FA hybrid algorithms, GA algorithms are used to build solution spaces, while FA is used to improve solutions. Although there was an increase in the fitness value of GFAKP after the sequential implementation of FA, the value remained below that of FAKP and FGAKP. In the Credit Approval dataset, it can be seen that FGAKP is 67.01% superior in terms of computing time to FAKP and 4.09% in terms of fitness value to GAKP. This is in line with Rovea [22] that states the FA-GA hybrid algorithm performs better in terms of computational time while maintaining the precision of the fitness value. As for the Heart Disease dataset, FGAKP is 14.76% superior in terms of computing time to FAKP and 8.46% in terms of fitness value to GAKP. Just as in the Acute Inflammations dataset, there was an increase in the fitness value of GFAKP after the sequential implementation of FA. However, the value was still below that of FAKP and FGAKP.

3.3 Comparison of Compactness Results among KP, FAKP, GAKP, and FGAKP on Benchmark Datasets

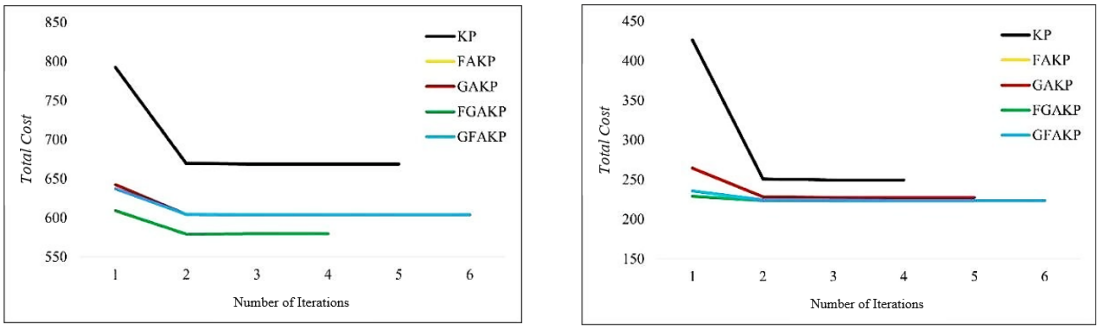
After obtaining the centroid initialization, grouping was carried out using the K-Prototype (KP) algorithm. At this stage, KP is run with two centroid initialization approaches: initialization of optimization results obtained from the FAKP, GAKP, FGAKP, and GFAKP algorithms, and random initialization without an optimization process. This process aims to evaluate the effectiveness of centroid initialization resulting from the metaheuristic optimization algorithm on all four test datasets. First of all, the quality of the KP grouping results without optimization and with optimization on the benchmark data is compared from the Total Cost (TC) per iteration, which is presented in Figure 3.4 as follows:



(1) Zoo



(2) Acute Inflammations



(3) Credit Approval (4) Heart Disease

Figure 3.4. Total Cost (TC) Change per Iteration of Benchmark Datasets

Based on Figure 3.4, it can be seen that the KP algorithm without optimization consistently produces a TC value that is higher in each iteration than the KP algorithm with optimization, across all trial data used in this study. This suggests that the randomly determined centroid initialization in the KP algorithm is not sufficiently effective in capturing the actual structure of the data distribution, resulting in a longer iteration process that tends to converge to poorly localized solutions.

Furthermore, the performance of the final results from each algorithm was evaluated using Total Cost (TC), Cluster Validity Index (CV), the number of iterations, and computational time. According to Table 3.2, FAKP, GAKP, FGAKP, and GFAKP consistently achieved lower TC values than KP without optimization across the research dataset. This means that the four optimization algorithms used can result in more centralized grouping at a lower cost. TC between FAKP and FGAKP has the same values. Here is a comparison table of TC on each algorithm:

Table 3.2. Comparison of Total Cost (TC) of Clustering Results on Benchmark Datasets

Dataset	KP	FAKP	GAKP	FGAKP	GFAKP
Zoo	90.0387	79.5856	80.3113	79.5856	80.3113
Acute Inflammations	48.7719	22.2516	22.3301	22.2516	28.2903
Credit Approval	668.8456	579.6264	603.7698	579.6264	603.7698
Heart Disease	249.3308	223.3046	227.3063	223.3046	223.5160

Based on Table 3.3, the FAKP, GAKP, FGAKP, and GFAKP algorithms have a larger CV index than KP without optimization in almost all research data. The increase in the CV index of the four optimization algorithms used indicated that the clusters formed through centroid initialization of the optimization method became more homogeneous for objects within the same cluster, as well as heterogeneous for objects from different clusters. From the results of the CV index evaluation, it can be seen that FAKP and FGAKP are the most optimal algorithms, as they consistently increase the CV index value the most across all trial data in this study. Here is a comparison table of CV indices on each algorithm:

Table 3.3. Comparison of Cluster Validity (CV) Index on Benchmark Datasets

Dataset	KP	FAKP	GAKP	FGAKP	GFAKP
Zoo	0.80140	0.88317	0.87017	0.88317	0.87017
Acute Inflammations	0.15984	0.31092	0.30644	0.31092	0.30831
Credit Approval	0.01333	0.01849	0.01666	0.01849	0.01666
Heart Disease	0.01866	0.01947	0.01835	0.01947	0.01915

Based on Table 3.4, in terms of the number of iterations (I) and the computational time in seconds (t) to achieve convergence, KP algorithms with optimization, especially FGAKP, require fewer iterations and shorter compute time to achieve convergence than KP without optimization in almost all of the research trial data. This means that the FGAKP algorithm finds stable solutions faster (in fewer iteration cycles), so that the search process is more targeted and does not require a lot of repair or iteration. Here is a comparison table of the number of iterations (I) and computational time in seconds (t) on each algorithm:

Table 3.4. Comparison of Iterations (I) and Computational Time (t) on Benchmark Datasets

Dataset	KP		FAKP		GAKP		FGAKP		GFAKP	
	I	***t***	***I***	***t***	***I***	***t***	***I***	***t***	***I***	***t***
Zoo	4	0.688	3	0.413	3	0.409	3	0.403	3	0.405
Acute Inflammations	3	0.480	2	0.133	2	0.127	2	0.120	3	0.286
Credit Approval	5	1.571	4	1.113	6	1.641	4	1.105	6	1.675
Heart Disease	4	0.938	5	0.979	5	0.975	5	0.973	6	1.149

Overall, based on the evaluation metrics—Total Cost (TC), Cluster Validity Index (CV), number of iterations (I), and computational time (t)—the comparison highlights that the optimized KP variants (FAKP, GAKP, FGAKP, and GFAKP) consistently outperform the standard KP without optimization across almost all evaluation metrics. This indicates that the initialization of the initial centroids has a significant impact on the clustering results, as better starting points lead to more optimal solutions, faster convergence, and improved cluster quality. In the main evaluation metrics, namely TC and CV, FAKP and FGAKP yield identical values. This is due to the iterative nature of the KP algorithm, which continues refining cluster assignments after generating the initial centroids. While FAKP and FGAKP may start from different centroids, the iterative process guides both toward the same optimal solution. Thus, the search spaces explored by FA and the FA–GA hybrid produce comparable high-quality starting points, ultimately leading to identical final centroids. However, a clearer difference appears in the centroid initialization stage (Section 3.2). FAKP requires a longer initialization time, while FGAKP reduces it by more than 60%. This shows that, despite achieving identical TC and CV, FGAKP is significantly more efficient in generating high-quality initial centroids, leading to fewer iterations and lower overall computational time. Therefore, FGAKP is the best clustering algorithm in this study.

3.4 Implementation of Village Development Clustering Based on BVI Indicators in North Kalimantan Province (2024) using the Best Proposed Algorithm

To evaluate the performance of the FGAKP algorithm on a real dataset, this study uses village development grouping. A comparative analysis is then conducted with the original algorithm, namely K-Prototype (KP), without optimization. Based on Figure 3.5, it can be observed that FGAKP consistently results in a lower Total Cost (TC) compared to KP without optimization over the entire iteration. During the first iteration, the TC value of the KP algorithm without optimization was very high, at 4285.136, while FGAKP was able to produce a TC 41.69% lower in the same iteration. The higher TC observed in the standard K-Prototype (KP) algorithm without optimization is primarily due to its random initialization of cluster centroids in the first iteration and its local search mechanism in subsequent iterations. These factors make the algorithm sensitive to initial conditions and prone to local optima. Consequently, the initial distances between data objects and centroids are relatively large, leading to a high TC value. In contrast, FGAKP significantly reduces TC by applying a hybrid metaheuristic approach in the centroid initialization phase. The Firefly Algorithm (FA) is first employed to intensify the search by guiding candidate solutions toward

regions with better fitness values, thereby producing more structured and less random initial solutions. Subsequently, the Genetic Algorithm (GA) enhances these solutions through exploration and refinement using crossover, mutation, and selection operators, enabling the algorithm to escape potential local optima and further improve solution quality. This sequential combination allows FGAKP to start from a near-optimal set of centroids, resulting in a substantially lower TC even in the first iteration compared to the KP algorithm without optimization. Building on this strong initialization, the algorithm achieves more stable and efficient reduction in TC throughout the subsequent iterations. When each algorithm converges, the use of FGAKP can reduce TC by 11.69% compared to KP without optimization. This significant difference indicates that FGAKP performs better in the grouping process, as it optimally minimizes TC in the grouping of village progress and independence conditions in North Kalimantan Province in 2024.

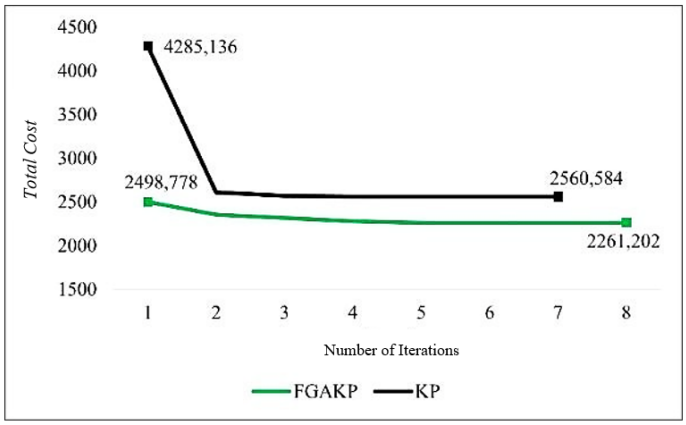


Figure 3.5. Comparison of Total Cost between FGAKP and Unoptimized KP for Village Development Clustering in North Kalimantan (2024)

Furthermore, the performance of village development groupings was analyzed from the evaluation metrics of the CV index, I, and t. Based on Table 3.5, it can be observed that the FGAKP algorithm produces a CV index of 8.916×10^{-3} or 16.75% greater than the KP without optimization. This higher CV index value indicates that FGAKP results in better grouping, as it shows a combination of homogeneity in categorical attributes and low numerical variance. This aligns with research by Hsu and Chen [6], indicating that the greater the value of the CV index, the better the grouping produced. In terms of the number of iterations, FGAKP requires eight iterations to converge, slightly higher than KP without optimization, which converges in seven iterations. Consequently, the computational time of FGAKP is also marginally longer. However, the increase in computational time (0.15 seconds) is negligible compared to the significant improvement in clustering quality. This indicates that the additional computational cost is justified by the more stable and optimal grouping results produced by FGAKP. Furthermore, the additional iteration in FGAKP reflects its ability to explore a broader solution space, leading to more representative and robust clusters. Therefore, considering both clustering quality and computational efficiency, FGAKP demonstrates superior performance in grouping village development conditions in North Kalimantan Province in 2024.

Table 3.5. Comparison of Cluster Validity (CV) Index, Iterations (I) and Computation Time (t) for Village Development Clustering

Dataset	CV Index	Number of Iterations (I)	Computation Time (t)
FGAKP	8.916×10^{-3}	8	4.776

KP	7.637×10^{-3}	7	4.625

After initializing the optimization results, the centroids of the village grouping will continue to be updated until they converge. This post-convergence centroid will be the final centroid in the grouping of conditions of village progress and independence in North Kalimantan Province in 2024. The final centroid of the village grouping with FGAKP shows some adjustment in the value distribution, but overall, the characteristic pattern remains similar to the initial initialization. This indicates that FGAKP can produce representative centroid initializations, resulting in a grouping process that is very close to the optimal solution, even from the initial iteration.

Based on the final centroid of FGAKP, each village is grouped according to the most similar characteristics within each cluster. Table VII shows the number and percentage of villages per cluster from the grouping results using the FGAKP algorithm. It can be observed that clusters 1 and 3 have the most significant proportion of villages, at 23.35%, while cluster 2 has the smallest proportion, at 10.74%. The following is a summary of the number and percentage of villages per cluster in the grouping of conditions of village progress and independence in North Kalimantan Province in 2024 with FGAKP:

Table 3.6. Distribution of Villages per Cluster using FGAKP Algorithm

Cluster	Number of Villages	Percentage
Cluster 1	113	23.35%
Cluster 2	52	10.74%
Cluster 3	113	23.35%
Cluster 4	100	20.66%
Cluster 5	106	21.90%

To determine the specific characteristics of each cluster, profiling was conducted on the clusters formed. The interpretation of the grouping results was reviewed from three aspects: social, economic, and environmental. Based on the analysis of the village grouping characteristics in North Kalimantan Province in 2024, in collaboration with FGAKP, the names for each cluster will be determined. The naming of village groups in each cluster is entirely an interpretation by the researcher based on the analysis of the characteristics of the grouping results. It therefore does not directly represent or determine the official status of the village as stipulated in the Building Village Index (BVI). The following is the naming of the village groups for each cluster, which is visually presented in the Figure 3.6:

1. Cluster 1 is labeled “Developing Village”, The villages in this group excel in fundamental aspects, including health, cooperatives, village-owned enterprises, the development of shopping groups, regional openness, road infrastructure, public transportation, and IT infrastructure. Cluster 1 villages can focus on improving security, particularly by constructing *kamling* posts and activating the making system. The government can also increase open public spaces, encourage the development of micro and small industries, and strengthen disaster mitigation programs.
2. Cluster 2 is labeled “Transitioning Village”. The villages in this group share characteristics similar to those in cluster 1. However, they still require more intensive development, particularly in basic services such as increasing the number of health workers and establishing health posts to improve access to health services. The government can focus on increasing economic activities, particularly the development of banks and BPRs, cooperatives, village-owned enterprises, and shop groups. In addition, villages in this cluster can make improvements in the construction of asphalt/concrete roads, increase public transportation, and educate the public so that they do not throw garbage into sewers/sewers.

3. Cluster 3 is labeled “Developed Village”. Villages in this group consistently show excellent social, economic, and environmental resilience. The government can focus on developing fixed routes on public transportation and encouraging the activation of the system to further strengthen the already sound security system.
4. Cluster 4 is labeled “Lagging Village”. Villages in this group have very inadequate resilience, both socially, economically, and environmentally. The government can prioritize basic areas, such as education, health, and the economy, which have a broad impact on the welfare conditions of villages and communities.
5. Cluster 5 is labeled “Self-Sustaining Village”. Villages in this group already possess good resilience in all aspects, but still require improvement in both quantity and quality.

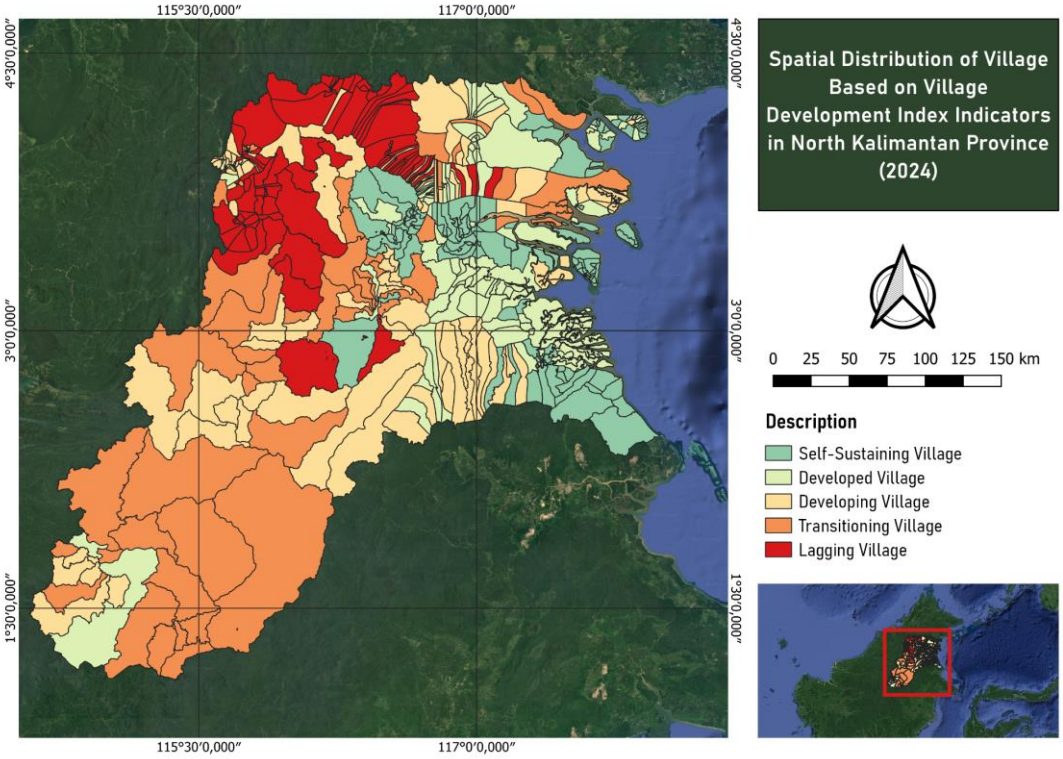


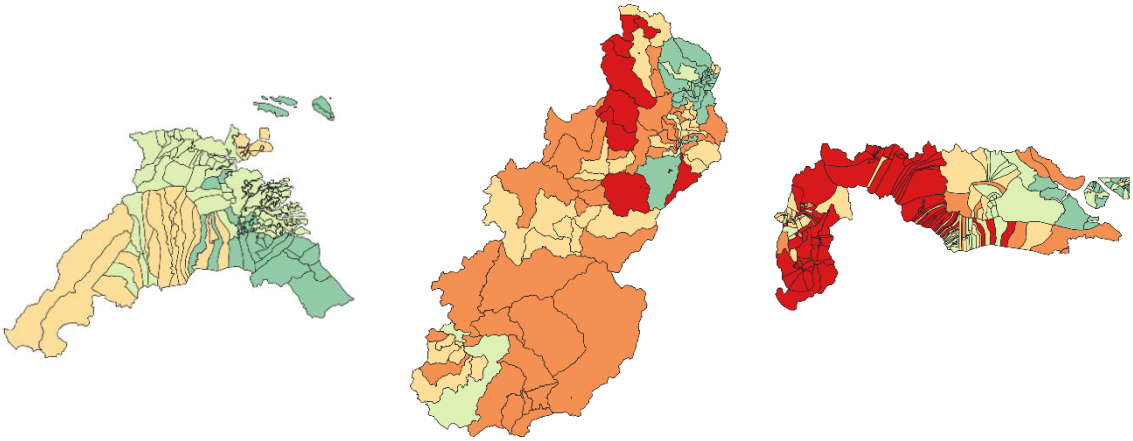
Figure 3.6. Spatial Distribution of Village Groups in North Kalimantan Based on BVI (2024)

To determine the conditions of progress and village independence more specifically, the results of the grouping were observed at the regency level. It can be observed that:

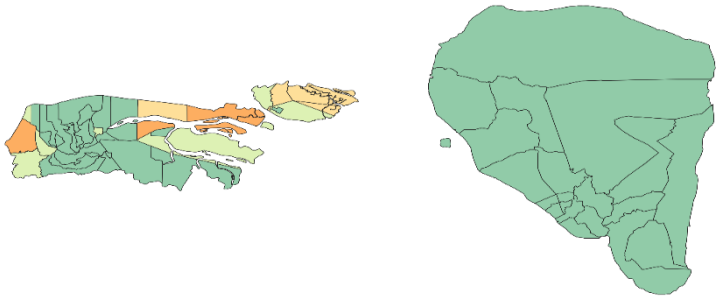
1. Developed and Self-Sustaining Village groups dominate by villages in Bulungan Regency. Additionally, there are no villages categorized as Lagging or Transitioning. This condition reflects the success of village development in Bulungan Regency, particularly in terms of meeting indicators of village independence and progress, including infrastructure, basic services, and the local economy. Furthermore, local governments can focus their interventions on developing villages, thereby accelerating their progress towards becoming developed or independent villages.
2. Transitioning and Developing Village groups dominate by villages in Malinau Regency. This means that most villages in Malinau Regency have a pretty good level of Development and are in a transition phase towards more optimal village development. Malinau Regency has an excellent opportunity to improve the welfare of its village community if targeted policies and interventions are implemented. In addition, the government also needs to map the specific

challenges faced by eight villages classified as Lagging Village, so that the interventions carried out are more targeted, especially in fulfilling basic needs services.

3. Lagging Village dominates by villages in Nunukan Regency. This shows that there are still many villages that face serious challenges in terms of basic infrastructure, public services, and local economic development. In addition, the proportion of Developing Village also dominates in Nunukan Regency. This means that most villages are still in the early or intermediate stages of village development, with a focus on improving the quality of life for their people. The government can focus on efforts to accelerate development in Lagging and Developing Village. An inclusive and locally driven, potential-based development approach is crucial for minimizing the gaps between villages and achieving more equitable progress in Nunukan Regency.
4. Developed and Self-Sustaining Village groups dominate by villages in Tana Tidung Regency. This indicates that the majority of villages in Tana Tidung Regency have successfully met various indicators of progress and village independence, encompassing social, economic, and environmental aspects. There are no villages categorized as Lagging, and only a small number of villages are categorized as Transitioning and Developing. This indicates that the equitable distribution of village development in Tana Tidung has been running well and relatively evenly.
5. All villages/sub-districts in Tarakan City are categorized as Self-Sustaining Village. There is not a single village that falls into the categories of Developed, Developing, Transitioning, or Lagging. This condition indicates that Tarakan City has a highly equitable and empowered village development, particularly in social, economic, and environmental aspects. The independence of villages in Tarakan also indicates the high quality of life of the community and the effectiveness of local development programs.



(1) Bulungan Regency (2) Malinau Regency (3) Nunukan Regency



(4) Tana Tidung Regency (5) Tarakan City

Figure 3.7. Thematic Maps of Village Development Status in North Kalimantan by Regency Based on Bvi Indicators (2024)

4. CONCLUSION

Based on the comprehensive analysis conducted on the benchmark datasets (Zoo, Acute Inflammations, Credit Approval, and Heart Disease), this study concludes that the proposed Hybrid Firefly-Genetic Algorithm K-Prototype (FGAKP) consistently outperforms the standard K-Prototype (KP), as well as the FAKP, GAKP, and GFAKP variants. On the other hand, the evaluation using Total Cost (TC) and Cluster Validity (CV) Index indicates FGAKP produced the most optimal cluster structures.

The implementation of FGAKP in grouping village development conditions in North Kalimantan Province, based on the Building Village Index (BVI) using Podes 2024 data, successfully reduces TC by 11.69% and increases CV by 16.75% compared to KP without optimization, resulting in five meaningful village clusters. These clusters provide a more adaptive representation of village conditions compared to the conventional BVI scoring approach, which applies uniform weights across all regions and may not fully capture local heterogeneity.

Cluster 1 is a group of villages that have sufficient development in terms of fundamental aspects, thereby being labeled as “Developing Village”. Cluster 2 is a group that shares similar characteristics with Cluster 1, but still lacks adequate access to basic services, and is therefore labeled as “Transitioning Village”. Cluster 3 is a group that consistently demonstrates excellent social, economic, and environmental resilience and is therefore labeled as “Developed Village”. Cluster 4 is a group with very inadequate conditions in almost all aspects of observation, and therefore, it is labeled as “Lagging Village”. Cluster 5 is a group with good resilience conditions in all aspects; however, it still needs to be improved in both quantity and quality to be labeled as “Self-Sustaining Village”.

The proposed approach offers a more flexible and data-driven alternative for identifying village characteristics, rather than relying on rigid scoring schemes. As a result, clustering results can support policymakers in designing more targeted, context-specific development strategies, improving resource allocation, and enhancing the accuracy of policy decision. Furthermore, the FGAKP framework can be generalized to other domains involving large-scale and heterogeneous data, where traditional scoring or rule-based approaches are insufficient. Thus, this study not only improves clustering performance but also contributes to more reliable and adaptive decision-making in real-world applications.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest

REFERENCES

[1] Abdullah, A., Deris, S., Mohamad, S., Zaiton, S., & Hashim, M., 2012. A New Hybrid Firefly Algorithm for Complex and Nonlinear Problem. *Advances in Intelligent and Soft Computing (AISC), 151,* 673–680. https://doi.org/10.1007/978-3-642-28765-7_81

[2] Ali, N., Othman, A., & Misran, M. H., 2014. A Review of Firefly Algorithm. *Asian Research Publishing Network (ARPN) Journal of Engineering and Applied Sciences, 9*(10), 1732–1736. www.arpnjournals.com

[3] Elkhechafi, M., Hachimi, H., & Elkettani, Y., 2017. A New Hybrid Firefly With Genetic Algorithm for Global Optimization. *International Journal of Management and Applied Science, 3*(4), 47–51. http://iraj

[4] Ezzeldin, R., Zelenakova, M., Abd-Elhamid, H. F., Pietrucha-Urbanik, K., & Elabd, S., 2023. Hybrid Optimization Algorithms of Firefly with GA and PSO for the Optimal Design of Water Distribution Networks. *Water (Switzerland), 15*(10), 1–15. https://doi.org/10.3390/w15101906

[5] Farahani, S. M., Abshouri, A. A., Nasiri, B., & Meybodi, M. R., 2012. Some Hybrid Models to Improve Firefly Algorithm Performance. *International Journal of Artificial Intelligence, 8*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=baf0337d48539f33150622a2bfbb89fc34cf93cd

[6] Hsu, C. C., & Chen, Y. C., 2007. Mining of Mixed Data with Application to Catalog Marketing. *Expert Systems with Applications, 32*(1), 12–23. https://doi.org/10.1016/j.eswa.2005.11.017

[7] Huang, T., Yin, H., & Huang, X., 2024. Improved Genetic Algorithm for Multi-Threshold Optimization in Digital Pathology Image Segmentation. *Sci Rep, 14*(22454), 1–21. https://doi.org/10.1038/s41598-024-73335-6

[8] Izamahendra, Y., 2025. Penerapan Metode Analisis Klaster dalam Pengelompokkan Desa di Kecamatan Pariaman Timur Berdasarkan Indeks Desa Membangun 2023. *Jurnal Informatika Dan Teknik Elektro Terapan, 13*(2), 204–211. https://doi.org/10.23960/jitet.v13i2.6138

[9] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R., 1988. Heart Disease [Dataset]. *UCI Machine Learning Repository*. https://doi.org/10.24432/C52P4X

[10] Kaur, A., Kumar, Y., & Sidhu, J., 2024. Exploring Meta-heuristics for Partitional Clustering: Methods , Metrics , Datasets , and Challenges. *Artificial Intelligence Review, 57*(10). https://doi.org/10.1007/s10462-024-10920-1

[11] Kemendesa PDTT., 2016. *Permendesa Nomor 2 Tahun 2016*. Diakses pada 22 November 2024, dari https://peraturan.bpk.go.id/Details/150585/permendes-pdtt-no-2-tahun-2016

[12] Kemendesa PDTT., 2024. *Peringkat Prov-Kab-Kec IDM 2024*. Diakses pada 2 Februari 2025, dari https://idm.kemendesa.go.id/

[13] Kemendesa PDTT., 2024. *SOP IDM 2024*. Diakses pada 13 Desember 2024, dari https://idm.kemendesa.go.id/view/detil/3/publikasi

[14] Nand, R., & Sharma, P., 2019. Iteration Split with Firefly Algorithm and Genetic Algorithm to Solve Multidimensional Knapsack Problems. *2019 IEEE Asia-Pacific Conference on Computer Science and Data Engineering, CSDE 2019*. https://doi.org/10.1109/CSDE48274.2019.9162422

[15] Nooraeni, R., 2015. Cluster Method Using A Combination of Cluster K-Prototype Algorithm and Genetic Algorithm for Mixed Data. *Jurnal Aplikasi Statistika & Komputasi Statistik, 7*(2), 81–97. https://doi.org/https://doi.org/10.34123/jurnalasks.v7i2.23

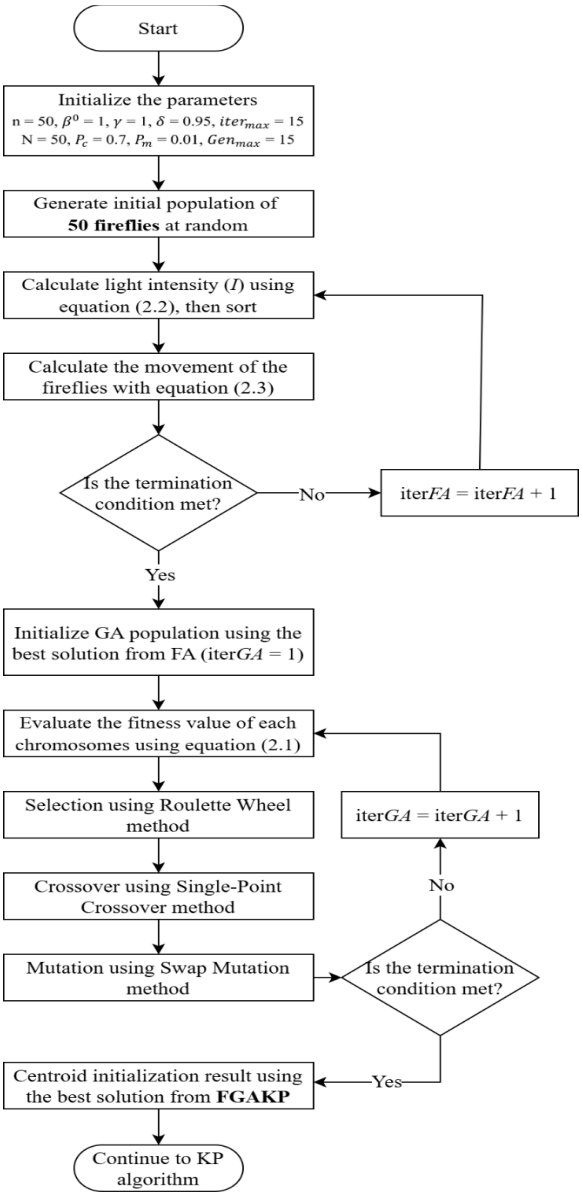
[16] Nooraeni, R., & Nurfalah, G., 2022. Kajian Penerapan Jarak Euclidean , Manhattan , Minkowski , dan Chebyshev pada Algoritma Clustering K-Prototype. *Sains, Aplikasi, Komputasi Dan Teknologi Informasi, 4*(2), 72–82. https://doi.org/10.30872/jsakti.v4i2.9241

[17] Nooraeni, R., Suprijadi, J., & Zulhanif., 2019. K-Prototype untuk Pengelompokan Sata Campuran. *Jurnal Statistika Teori dan Aplikasi: Biomedics, Industry & Business and Social Statistics*, *13*(1), 9–16. https://www.researchgate.net/publication/337847692
[18] Okwu, M. O., & Tartibu, L. K., 2021. *Metaheuristic Optimization: Nature-Inspired Algorithms Swarm and Computational Intelligence, Theory and Applications* (Vol. 927). https://doi.org/https://doi.org/10.1007/978-3-030-61111-8
[19] Pham, D. -T., Suarez-Alvarez, M. M., & Prostov , Y. I., 2011. Random Search with K -Prototypes Algorithm for Clustering Mixed Datasets. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, *467*(2132), 2387–2403. https://doi.org/10.1098/rspa.2010.0594
[20] Radhwani, A. M. N. A. & Algamal, Z. Y., 2021. Improving K-Means Clustering based on Firefly Algorithm. *Journal of Physics: Conference Series*, *1897*(12004), 1–8 https://doi.org/10.1088/1742-6596/1897/1/012004
[21] Revata, S., & Rino., 2023. Implementation of Data Mining Classification of People ' s Personalities using Naïve Bayes Algorithm. *Bit-Tech*, *6*(1). https://doi.org/10.32877/bt.v6i1.741
[22] Rovea, O., 2014. Genetic algorithm and firefly algorithm hybrid schemes for cultivation processes modelling. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8790*, 1–4. https://doi.org/10.1007/978-3-662-44994-3
[23] Samita, G. R., Wisesa, W., Setiawan, E. D., Respati K. I., & Hidayat, R., 2024. Integrasi Artificial Intelligence dan Teori Bounded Rationality dalam Mengatasi Ketidakpastian Pengambilan Keputusan Bisnis di Era Big Data. *Jurnal Bisnis Dan Komunikasi Digital*, *2*(2), 1–12. https://doi.org/10.47134/jbkd.v2i2.3460
[24] Senthilnath, J., Omkar, S. N., & Mani, V., 2011. Clustering using firefly algorithm: Performance study. *Swarm and Evolutionary Computation*, *1*(3), 164–171. https://doi.org/10.1016/j.swevo.2011.06.003
[25] Sihotang, H. T., Albert, M., Riandari, F., & Rendell, L., 2023. Efficient Optimization Algorithms for Various Machine Learning Tasks, Including Classification, Regression, and Clustering. *Idea : Future Research*, *1*(1). https://doi.org/http://dx.doi.org/10.35335/idea.v1i1.3
[26] Song, J., Wang, B., & Hao, X., 2024. Optimization Algorithms and Their Applications and Prospects in Manufacturing Engineering. *Materials*, *17*(16). https://doi.org/10.3390/ma17164093
[27] Suwirmayanti, P., Putra, I. K. G. D., & Kumara, I. N. S., 2014. Optimasi pusat cluster K-Prototype dengan algoritma genetika. *Majalah Ilmiah Teknologi Elektro*, *13*(2). https://doi.org/10.24843/10.24843/MITE.2015.v13i02p02
[28] Wang, L., Duan, H., Liu, Z., Peng, Y., & Liu, X., 2024. Research on Modeling Method for Optimal Allocation of Wellhead Targets in Large Well Clusters. *Processes*, *12*(8), 1–11. https://doi.org/10.3390/pr12081705
[29] Yang, X. -S., 2013. *Cuckoo Search and Firefly Algorithm*. https://doi.org/https://doi.org/10.1007/978-3-319-02141-6
[30] Yang, X. -S., & He, X., 2013. Firefly Algorithm : Recent Advances and Applications. *Int. J. Swarm Intelligence*, *1*(1), 36–50. https://doi.org/https://doi.org/10.1504/IJSI.2013.055801
[31] Yang, X. -S., 2010. *Nature-Inspired Metaheuristic Algorithms Second Edition*. Luniver Press. https://staff.fmi.uvt.ro/~daniela.zaharie/ma2016/projects/techniques/FireflyAlgorithm/Yang_nature_book_part.pdf
[32] Yuslimah., 2023. Peran Big Data terhadap Kualitas Data Statistik Indonesia di Era Digital. *Madani: Jurnal Ilmiah Multidisiplin*, *1*(11). https://doi.org/10.5281/zenodo.10402327.

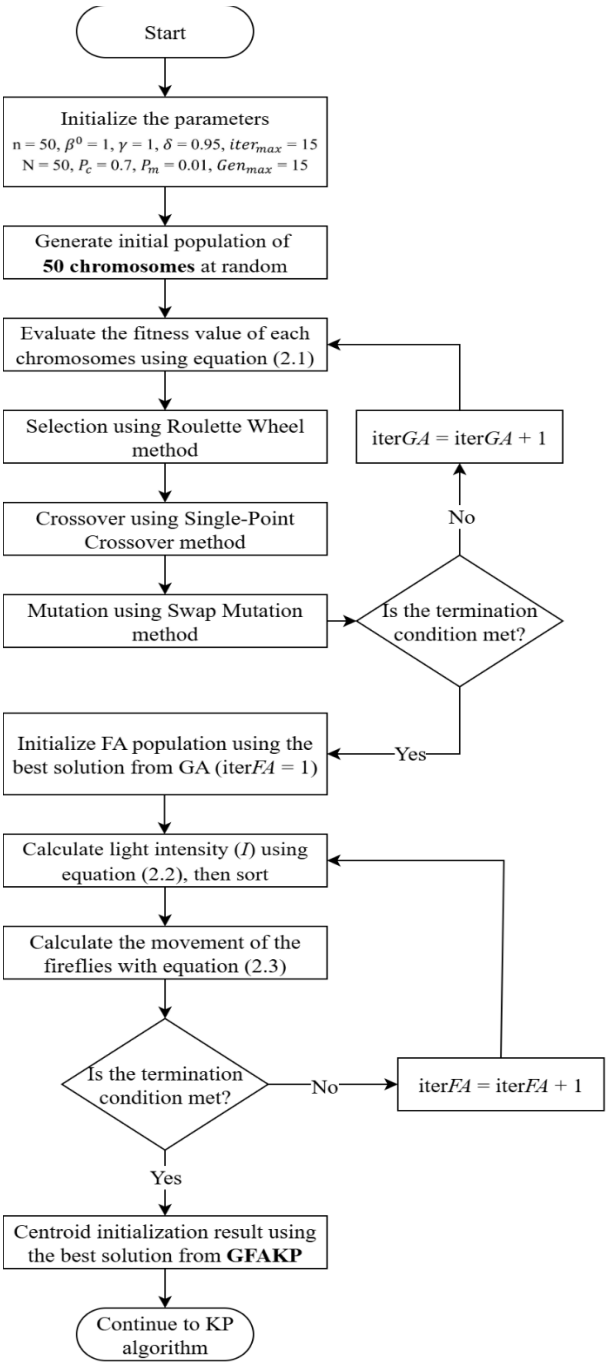
[33] Zhang, W., Jiao, C., & Zhou, Q., 2025. Firefly Algorithm with Multiple Learning Ability based On Gender Difference. *Sci Rep, 15*(28400), 1–31. https://doi.org/10.1038/s41598-025-09523-9

APPENDICES

Appendix 1. Flowchart of the FGAKP Algorithm



Appendix 2. Flowchart of the GFAKP Algorithm


Start
Initialize the parameters
n = 50, $\beta^0 = 1$, $\gamma = 1$, $\delta = 0.95$, $iter_{max} = 15$
N = 50, $P_c = 0.7$, $P_m = 0.01$, $Gen_{max} = 15$
Generate initial population of 50 chromosomes at random
Evaluate the fitness value of each chromosomes using equation (2.1)
Selection using Roulette Wheel method
Crossover using Single-Point Crossover method
Mutation using Swap Mutation method
Is the termination condition met?
No
iterGA = iterGA + 1
Yes
Initialize FA population using the best solution from GA (iterFA = 1)
Calculate light intensity (I) using equation (2.2), then sort
Calculate the movement of the fireflies with equation (2.3)
Is the termination condition met?
No
iterFA = iterFA + 1
Yes
Centroid initialization result using the best solution from GFAKP
Continue to KP algorithm