jowua paper

10859-29829-1-RV

icst

Document Details

Submission ID
trn:oid:::3618:117590528

Submission Date
20 Oct 2025, 13:51 GMT+7

Download Date
20 Oct 2025, 13:53 GMT+7

File Name
10859-29829-1-RV.docx

File Size
360.9 KB

8 Pages

4,204 Words

26,589 Characters

18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography

Match Groups

62 Not Cited or Quoted 16%
Matches with neither in-text citation nor quotation marks

7 Missing Quotations 2%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

7% Internet sources
8% Publications
14% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **62** Not Cited or Quoted 16%
  Matches with neither in-text citation nor quotation marks
- **7** Missing Quotations 2%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

7% Internet sources
8% Publications
14% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Student papers	
Universitas Prima Indonesia on 2025-06-14		2%
2	Student papers	
Liverpool John Moores University on 2024-06-17		2%
3	Publication	
John R. Owen, Ian T. Nabney, José L. Medina-Franco, Fabian López-Vallejo. "Visuali...		1%
4	Internet	
link.springer.com		<1%
5	Student papers	
Georgia Institute of Technology Main Campus on 2025-06-29		<1%
6	Student papers	
The Independent Institute of Education (IIE) on 2025-06-13		<1%
7	Student papers	
University of Queensland on 2025-05-30		<1%
8	Publication	
Maria Luiza Mitu, Dorina Ticoş, Nicoleta Udrea, Adrian Scurtu, Cătălin Mihai Ticoş....		<1%
9	Internet	
scienceofbiogenetics.com		<1%
10	Internet	
ejournal.ikado.ac.id		<1%

11 Student papers
Georgia Institute of Technology Main Campus on 2025-02-10 <1%

12 Student papers
Universitas Khairun on 2024-04-18 <1%

13 Internet
fastercapital.com <1%

14 Internet
jdmdc.com <1%

15 Student papers
2U Southern Methodist University on 2022-02-16 <1%

16 Internet
escape35-belgium.eu <1%

17 Student papers
Georgia Institute of Technology Main Campus on 2024-09-22 <1%

18 Publication
Thangaprakash Sengodan, Sanjay Misra, M Murugappan. "Advances in Electrical ... <1%

19 Student papers
RMIT University on 2024-10-13 <1%

20 Student papers
ICL Education Group on 2025-08-17 <1%

21 Publication
Onur Dogan, Hunaıda Avvad. "Fuzzy Clustering Based on Activity Sequence and C... <1%

22 Publication
Pasqual Martí, Jaume Jordán, Javier Palanca, Vicente Julian. "Charging stations an... <1%

23 Student papers
University of East London on 2025-05-02 <1%

24 Publication
Amin Golzari Oskouei, Negin Samadi, Shirin Khezri, Arezou Najafi Moghaddam et ... <1%

25 Student papers
Dublin City University on 2025-07-29 <1%

26 Student papers
Northcentral on 2025-09-28 <1%

27 Student papers
Nottingham Trent University on 2025-04-02 <1%

28 Student papers
Swinburne University of Technology on 2024-12-09 <1%

29 Internet
github.com <1%

30 Student papers
Coventry University on 2025-08-15 <1%

31 Student papers
Dublin Business School on 2025-08-28 <1%

32 Student papers
University of Stirling on 2025-04-11 <1%

33 Internet
ejurnal.seminar-id.com <1%

34 Internet
journals.uob.edu.bh <1%

35 Internet
peerj.com <1%

36 Internet
thesai.org <1%

37 Internet
www.mdpi.com <1%

38 Student papers
Leiden University on 2023-08-25 <1%

39 Student papers

University of Nicosia on 2024-12-30 <1%

40 Student papers

Glyndwr University on 2025-08-04 <1%

41 Publication

Hyuk-Gyu Park, Kwang-Seong Shin, Jong-Chan Kim. "Efficient Clustering Method f... <1%

42 Publication

Lei Shu, Guirong Li. "Application of improved clustering algorithm in mixed teach... <1%

43 Student papers

The Robert Gordon University on 2025-08-15 <1%

44 Student papers

Napier University on 2023-04-17 <1%

JIKO (Jurnal Informatika dan Komputer) Accredited KEMDIKTISAINTEK, No.0173/C3/DT.05.00/2025
Vol. x, No. x, April 2025, pp. x-x p-ISSN: 2614-8897
DOI: 10.33387/jiko e-ISSN: 2656-1948

PERFORMANCE EVALUATION OF HYBRID CLUSTERING K-MEANS AND DBSCAN WITH FEATURE WEIGHT OPTIMIZATION

Vic Devlin[1], Robet[2*], Octara Pribadi[3]

[1,2,3] Informatics Engineering, STMIK TIME, Medan, Indonesia
Email: [1]vicdevlin2004@gmail.com, [*2]robertdetime@gmail.com, [3]octarapribadi@gmail.com

(Received: dd mmm yyyy, Revised: dd mmm yyyy, Accepted: dd mmm yyyy)

Abstract

This research evaluates the performance of a hybrid clustering model that integrates K-Means and DBSCAN, enhanced through Feature Weight Optimization (FWO) using a Genetic Algorithm (GA), to achieve more precise consumer data segmentation. Two benchmark datasets, Customer Personality Analysis (CPA) and Online Retail (OR), were utilized to examine how different clustering techniques respond to variations in data structure. The feature weighting process was optimized using GA to improve the representational contribution of each variable toward the final cluster configuration. The Silhouette Score was adopted as the primary evaluation metric to measure intra-cluster cohesion and inter-cluster separation. Experimental findings reveal that for the CPA dataset, the Hybrid + FWO method achieved the best performance with a Silhouette Score of 0.9600, while the K-Means + FWO method recorded the highest score of 0.9804 on the OR dataset. Across all scenarios, the inclusion of FWO consistently enhanced clustering stability and interpretability. These results highlight that algorithm selection must consider dataset characteristics, and that feature weight optimization is pivotal in strengthening segmentation quality and ensuring more meaningful insights in consumer behavior analytics.

Keywords: *K-Means, DBSCAN, Hybrid Clustering, Feature Weight Optimization, Silhouette Score*

**Corresponding Author: Author2*

1. INTRODUCTION

The acceleration of digital transformation over the past two decades has fundamentally changed how organizations manage information and make strategic decisions. Data is now regarded as a valuable asset that can provide a competitive advantage when systematically managed and analytically processed [1]. In the modern business landscape, particularly in the retail and e-commerce sectors, the utilization of customer data has become crucial for understanding market needs and developing relevant marketing strategies. One effective analytical approach for leveraging such data is customer segmentation, which involves dividing consumers into groups based on similarities in attributes, transactional behavior, or preference patterns. Through segmentation, organizations can prioritize marketing strategies, deliver targeted promotions, and enhance customer loyalty [2]. Data-driven segmentation strategies have also proven effective in optimizing customer retention and minimizing marketing campaign costs [1], [2].

In the realm of unsupervised learning, clustering has emerged as one of the most widely used techniques to identify hidden patterns within unlabeled data. Among the various algorithms, K-Means clustering occupies a significant position due to its ability to efficiently partition data based on Euclidean distance from centroid points [3]. Its main advantages include computational efficiency and ease of implementation across analytical platforms. However, K-Means also exhibits notable limitations, such as sensitivity to initial centroid placement, failure to detect clusters of arbitrary shapes, and weakness in handling outliers [4]. To overcome these drawbacks, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was developed. DBSCAN groups data points based on density, offering the advantage of not requiring a predefined number of clusters and automatically detecting noise [5]. Recent

studies have introduced variants such as Adaptive Multi-Density DBSCAN (AMD-DBSCAN), which adapts to varying data densities, particularly in large-scale datasets [4].

As a further advancement, several studies have proposed hybrid clustering approaches that combine two or more algorithms to enhance the accuracy and stability of segmentation results. One of the most commonly adopted combinations is K-Means and DBSCAN, as these two algorithms complement each other: K-Means excels in computational efficiency, while DBSCAN performs better in identifying irregularly shaped clusters and handling noise [2], [5]. The hybrid approach produces more consistent and adaptive outcomes than individual methods, making it especially valuable in consumer segmentation tasks such as identifying loyal, potential, and inactive customers [1], [2]. However, existing studies rarely address how feature weighting and hybrid clustering can be integrated to improve cluster quality in large and heterogeneous consumer datasets. This gap highlights the need for a systematic approach that jointly optimizes clustering structure and feature contribution to achieve higher segmentation accuracy.

Beyond algorithm selection, another factor that significantly affects clustering performance is the relative contribution of each feature within the dataset. Unequal feature weights can degrade clustering accuracy by disproportionately influencing the formation of certain centroids. To mitigate this, Feature Weight Optimization (FWO) techniques are applied to adjust the weight of each attribute according to its significance within the data structure [6]. A widely adopted approach for this task is the Genetic Algorithm (GA), an optimization method inspired by the process of natural evolution that uses selection, mutation, and crossover to find near-optimal solutions. GA is known for its ability to explore large solution spaces and identify optimal feature weight combinations without assuming linearity [7]. Integrating GA into clustering has been shown to improve performance, as evidenced by higher evaluation metrics such as the Silhouette Score, which measures intra-cluster cohesion and inter-cluster separation [8]. Therefore, the use of GA for feature weight optimization represents an effective means of enhancing clustering accuracy, particularly for high-dimensional datasets.

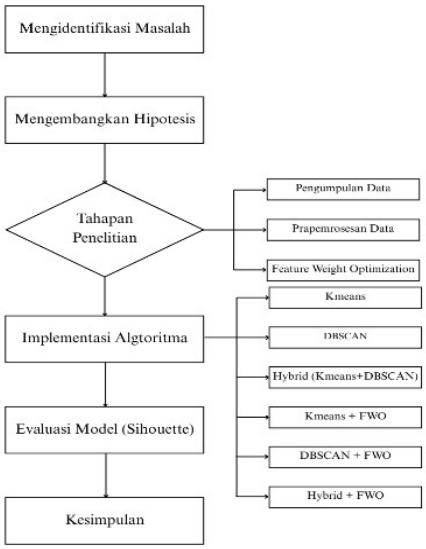
Clustering performance is typically evaluated using internal metrics such as the Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index. Among these, the Silhouette Score is the most widely adopted due to its intuitive representation of cluster quality, ranging from –1 to 1 [9]. A value close to 1 indicates that data points are well assigned to their respective clusters with strong separation. In large-scale data analysis, the Silhouette Score has been adapted to distributed computing environments such as Hadoop and Spark for efficient evaluation [9]. Principal Component Analysis (PCA) is also often employed to visualize clustering results, facilitating easier interpretation. Therefore, this study aims to evaluate the performance of a hybrid clustering approach that combines K-Means and DBSCAN with Feature Weight Optimization based on a Genetic Algorithm. The evaluation is conducted using the Silhouette Score as the primary metric and PCA for visual validation. The findings are expected to contribute to the development of more accurate, adaptive, and efficient customer segmentation methods for large-scale data environments.

2. RESEARCH METHOD

This research was designed using a quantitative experimental approach aimed at analyzing and evaluating the performance of a hybrid clustering algorithm. The proposed method integrates two widely used clustering techniques, K-Means and DBSCAN, which are subsequently enhanced through Feature Weight Optimization (FWO) using a Genetic Algorithm (GA) to improve customer segmentation accuracy. The overall workflow of this study includes problem identification, data collection, preprocessing, clustering implementation, feature weight optimization, and model evaluation.

Each stage of the research was structured according to the methodological frameworks proposed by Sridevi and Rajanna[5] as well as Kouser et al. [6]. These studies demonstrated that integrating partition-based methods (such as K-Means) with density-based approaches (such as DBSCAN) can yield more stable and adaptive cluster structures, especially when dealing with datasets characterized by varying data densities and heterogeneous feature distributions.

The research process was visualized as a flow diagram, as illustrated in Figure 1, which presents the sequential relationship among the major stages starting from data input and preprocessing to clustering execution, feature optimization, and model evaluation.



This diagram provides a comprehensive overview of the hybrid clustering process and its optimization

mechanism, serving as a conceptual foundation for the experimental procedures conducted in this study.

Figure 1. Research Flowchart

2.1 Dataset Description

This study utilized two publicly available datasets relevant to customer segmentation experiments, namely the Marketing Campaign Dataset and the Online Retail Dataset. These datasets were selected to represent two distinct but complementary consumer contexts behavioral marketing and transactional purchasing patterns thereby providing a valid foundation for cross-domain evaluation [2], [10].

1. Marketing Campaign Dataset, this dataset contains 2,240 customer records with a total of 29 attributes, encompassing demographic data (Year_Birth, Education, Marital_Status), economic variables (Income), and consumption behavior features such as MntWines, MntMeatProducts, and NumWebPurchases. The dataset was selected because it reflects real-world consumer marketing behavior, making it suitable for behavior-based segmentation analysis. The data are publicly available on the Kaggle platform [1].
2. Online Retail Dataset, the second dataset consists of e-commerce transaction records from a UK-based retail company, containing 541,909 entries and eight key attributes, namely InvoiceNo, StockCode, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. Since the data are transactional, aggregation was performed based on CustomerID to construct unique customer profiles. The aggregated features include:

a. TotalQuantity: The total number of items purchased per customer.
b. Frequency: The total number of unique transactions (InvoiceNo).
c. TotalSpent: The total expenditure of each customer, calculated using Equation (1).

$$TotalSpent = \sum_{i=1}^{n} (Quantity_i \times UnitPrice_i) \quad (1)$$

Both datasets were chosen because they represent two distinct perspectives of consumer behavior marketing interaction and real purchase transactions enabling a comprehensive evaluation of the proposed hybrid clustering model across different data characteristics [2], [10].

2.2 Data Preprocessing

The preprocessing stage was performed to ensure the integrity, consistency, and analytical readiness of the datasets before applying clustering algorithms. This stage consisted of three major steps: data cleaning, categorical transformation, and feature normalization.

1. Data Cleaning, records containing missing values and duplicate entries were removed to prevent bias and distortion in subsequent analysis. This process ensured that all remaining records were valid and representative of the dataset's true distribution.
2. Categorical Attribute Transformation, Categorical variables were converted into numerical representations using two techniques: Label Encoding for ordinal variables and One-Hot Encoding for nominal variables. This transformation allowed the clustering algorithms to process the data numerically while preserving categorical distinctions [11].
3. Feature Normalization, all numerical attributes were normalized within the range of [0, 1] using the Min–Max Scaling technique. This step was conducted to prevent attributes with larger magnitudes from dominating the distance calculations during clustering and to ensure fair contribution of each feature to the clustering process.

2.3 Clustering Method Implementation

1. K-means algorithm, the K-Means algorithm was employed to generate the initial customer segmentation by minimizing the Euclidean distance between data points and their respective centroids. The process begins by defining the number of clusters (k), followed by iterative centroid updates until the Sum of Squared Errors (SSE) objective function reaches a minimum value [3], [12]. For both datasets, the number of clusters was set to two (n_clusters = 2), aligning with the initial assumption of binary customer segmentation. The parameter random_state = 42 was applied to ensure reproducibility and consistency across multiple runs. This configuration was selected to provide a balanced baseline for comparative evaluation with other clustering approaches.
2. DBSCAN algorithm, The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm was used to identify cluster structures based on data density. Two key parameters were configured epsilon (ε), representing the neighborhood radius, and min_samples (MinPts), indicating the minimum number of points required to form a dense region [4]. The parameter tuning for DBSCAN was conducted separately for each dataset through exploratory testing, seeking combinations that achieved the highest Silhouette Score with a valid number of clusters (excluding the noise label "–1").

a. For the Online Retail dataset, the optimal configuration was found at (eps = 1.0, min_samples = 3), producing five clusters (including noise). Although this configuration achieved a Silhouette Score of 0.9400, the score was partially affected by the presence of outliers labeled as noise. Nevertheless, the visual distribution of the resulting clusters indicated

representative separability, validating the choice of parameters.

b. For the Customer Personality Analysis (CPA) dataset, the optimal parameters were (eps = 0.7, min_samples = 10), generating three valid clusters (including noise). This configuration effectively separated customer groups based on income and total expenditure, despite the existence of some noise points.

These parameter selections provided an optimal balance between inter-cluster separation and noise reduction, aligning with the recommendations of recent clustering optimization studies [13], [14].

3. Hybrid-clustering (K-Means + DBSCAN), the hybrid clustering approach was designed to combine the complementary strengths of K-Means and DBSCAN leveraging DBSCAN's ability to detect noise and K-Means' efficiency in handling spherical cluster structures. The implementation procedure was defined as follows:

a. Noise relabeling: For data points labeled "–1" (noise) by DBSCAN, their labels were reassigned using the predicted cluster assignments from K-Means.

b. Final labeling: The final cluster labels were derived from the DBSCAN output, with revised labels applied to former noise points as determined by K-Means.

c. Stability enhancement: This integration reduced the adverse effects of excessive noise while maintaining the structural consistency of the cluster formation.

Both datasets used the same parameter configurations as defined in the respective K-Means and DBSCAN implementations. The hybrid strategy improved cluster stability and minimized information loss caused by noise overrepresentation [15], [16].

2.4 Feature Weight Optimization (FWO)

The adjustment of feature weights in this study was performed through a Genetic Algorithm (GA), which is widely recognized for its ability to explore complex and high-dimensional search spaces while avoiding local optima. The primary objective of this optimization process was to adaptively tune the relative contribution of each feature to the final clustering structure.

Each individual in the GA population was represented as a weight vector whose dimension corresponded to the number of features used in the dataset. The overall procedure consisted of several key phases, as described below:

1. Initialization, an initial population of 20 individuals was generated, where each individual contained randomly assigned weights within the range of [0.1, 1.0]. These values served as the initial solution set to be evolved through successive generations.
2. Fitness evaluation, each individual (i.e., weight vector w) was evaluated by computing the Silhouette Score obtained from the clustering results. The fitness function was defined as:

$$f(w) = Silhouette\ (X \times w, labels) \quad (2)$$

where X denotes the normalized dataset, w represents the weight vector applied to each feature, and labels are the cluster assignments produced by K-Means, DBSCAN, or the Hybrid model.

The fitness function evaluates how well the clustering structure achieves both intra-cluster cohesion and inter-cluster separation. A higher fitness value indicates better clustering quality and more optimal feature weighting.

3. Selection, the Tournament Selection method was applied to identify individuals with superior fitness values. The best-performing individuals were chosen to propagate to the next generation, ensuring the survival of optimal solutions while maintaining population diversity.
4. Mutation, to introduce variability and prevent premature convergence, random perturbations were applied to individual weights using Gaussian Mutation, defined as:

$$w_i' = w_i + \mathcal{N}(0, \sigma^2) \quad (3)$$

where w_i is the original feature weight, w_i' is the mutated weight, and $N(0, \sigma^2)$ represents a Gaussian random variable with mean 0 and variance σ^2. This mechanism allows small random adjustments in the feature space, promoting exploration and maintaining diversity within the population.

5. Crossover, for the Online Retail Dataset, a Two-Point Crossover strategy was implemented to enhance solution variety. Two parent individuals exchange partial segments of their chromosomes between crossover points P_1 and P_2, producing new offspring. The crossover operation can be expressed as:

$$w_i' = \begin{cases} w_i^{(1)}, & i \in [\mathcal{P}_1, \mathcal{P}_2] \\ w_i^{(2)}, & otherwise \end{cases} \quad (4)$$

where $w_i^{(1)}$ and $w_i^{(2)}$ are the weight genes of two parent individuals, $\mathcal{P}_1$ and $\mathcal{P}_2$ are crossover points. This mechanism was selectively applied to the Online Retail dataset because it provided improved exploration and variation in the generated feature weight combinations, particularly effective for transactional data structures.

6. Evolutionary iteration, The GA evolutionary process consisting of selection, mutation, crossover, and fitness evaluation was iteratively executed over multiple generations until convergence or until the optimal weight configuration maximizing the fitness function was achieved. The parameter settings used in this study were as follows:
   a. Population size: 20
   b. Generations: 10
   c. Crossover probability (cxpb): 0.5
   d. Mutation probability (mutpb): 0.2
   e. Mutation type: Gaussian Mutation
   f. Selection: Tournament Selection (size = 3)

   The optimization process aimed to maximize the Silhouette-based objective function, enabling the model to dynamically adapt feature contributions for improved inter-cluster separation and intra-cluster cohesion. This confirmed the effectiveness of the Genetic Algorithm in enhancing clustering performance, particularly for high-dimensional datasets [6], [17].

2.4 Experimental Design of Clustering Methods

This study evaluated six clustering scenarios divided into two main categories: methods without feature weight optimization, and methods with Feature Weight Optimization (FWO) based on the Genetic Algorithm (GA). Each clustering technique was tested both independently and in a hybrid configuration, as summarized in Table 1.

Table 1. Experimental Scenarios

Clustering Method	Feature Weight Optimization (FWO)
K-Means	Not Applied
DBSCAN	Not Applied
Hybrid	Not Applied
K-Means	Applied
DBSCAN	Applied
Hybrid	Applied

All experiments were evaluated using the Silhouette Score metric, which measures intra-cluster cohesion and inter-cluster separation [9]. To complement quantitative evaluation, the clustering results were visualized using Principal Component Analysis (PCA), which projected the multi-dimensional data into two dimensions. This visualization provided a clearer understanding of the spatial distribution of clusters, aiding in the interpretation of how effectively each method separated distinct data groups [7], [18].

3. RESULT AND DISCUSSION

This section presents the experimental results of three clustering approaches K-Means, DBSCAN, and the Hybrid method (K-Means + DBSCAN) evaluated both without and with Feature Weight Optimization (FWO) using the Genetic Algorithm (GA). The performance of each method was assessed using the Silhouette Score, while the Principal Component Analysis (PCA) visualization was employed to illustrate the spatial distribution and cohesion of clusters.

3. 1 Customer Personality Analysis (CPA) Dataset

The experimental results for the Customer Personality Analysis (CPA) dataset are summarized in Table 2.

Table 2. Silhouette Score for CPA dataset

Clustering Method	FWO	Silhouette Score
K-Means	Not Applied	0.5951
DBSCAN	Not Applied	0.7478
Hybrid	Not Applied	0.7614
K-Means	Applied	0.9592
DBSCAN	Applied	0.8979
Hybrid	Applied	0.9599

The PCA projections for all six experimental scenarios are shown in Figures 2–7, illustrating how each algorithm separated the consumer data.

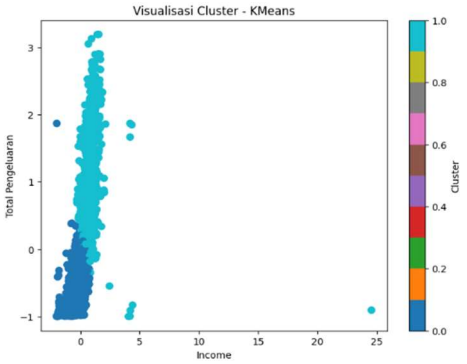


Figure 2. PCA Visualization of K-Means (CPA)

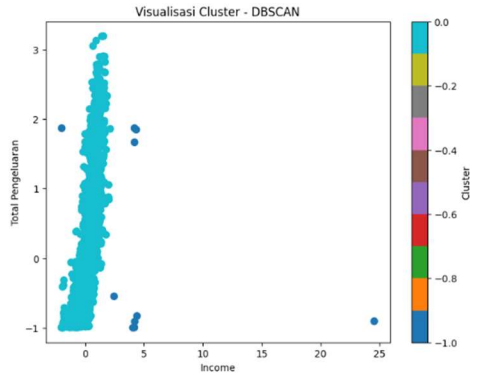


Figure 3. PCA Visualization of DBSCAN (CPA)

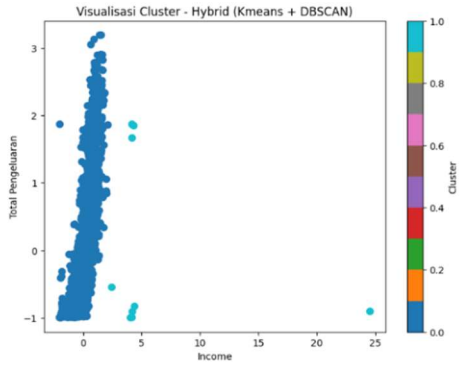


Figure 4. PCA Visualization of Hybrid Method (CPA)

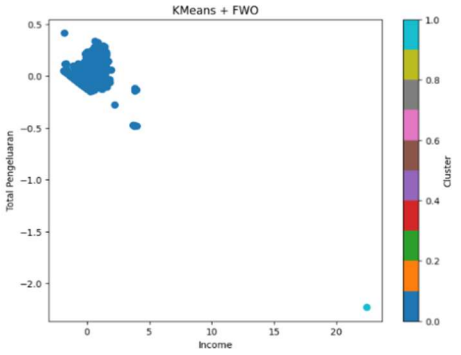


Figure 5. PCA Visualization of K-Means + FWO (CPA)

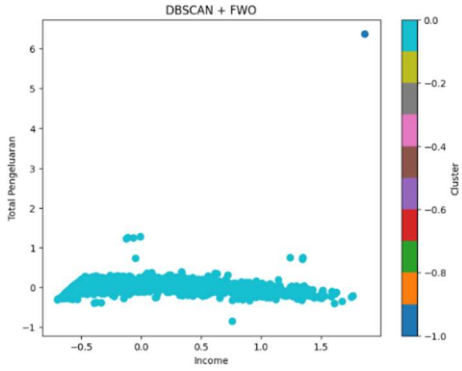


Figure 6. PCA Visualization of DBSCAN + FWO (CPA)

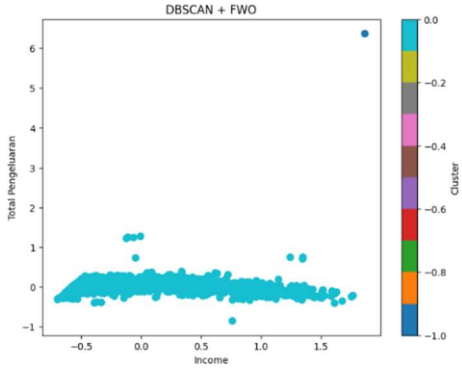


Figure 7. PCA Visualization of Hybrid Method +FWO (CPA)

The distribution in Figure 7 exhibits the most optimal separation, with minimal noise and well-defined cluster boundaries. This result demonstrates that integrating FWO significantly enhances the hybrid model's ability to form compact and well-separated clusters.

3. 2 Online Retail (OR) Dataset

The experimental results for the Online Retail (OR) dataset are summarized in Table 3.

Table 3. Silhouette Score for OR dataset

Clustering Method	FWO	Silhouette Score
K-Means	Not Applied	0.9801
DBSCAN	Not Applied	0.9400
Hybrid	Not Applied	0.9452
K-Means	Applied	0.9804
DBSCAN	Applied	0.9435
Hybrid	Applied	0.9521

The PCA projections for all six experimental scenarios are shown in Figures 8–13, illustrating how each algorithm separated the consumer data.

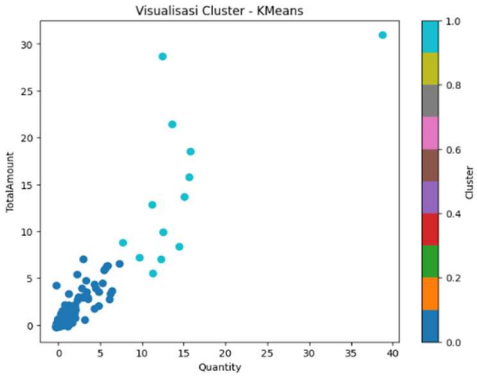


Figure 8. PCA Visualization of K-Means (OR)

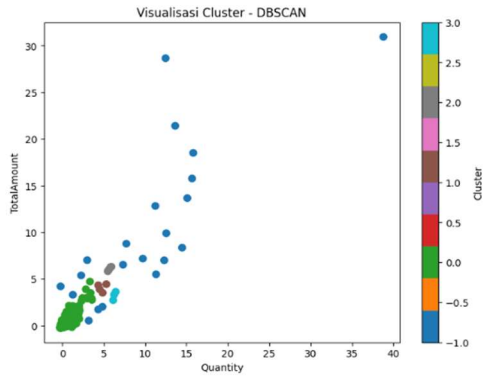


Figure 9. PCA Visualization of DBSCAN (OR)

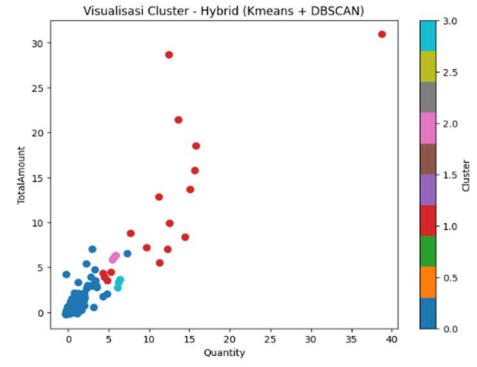


Figure 10. PCA Visualization of Hybrid Method (OR)

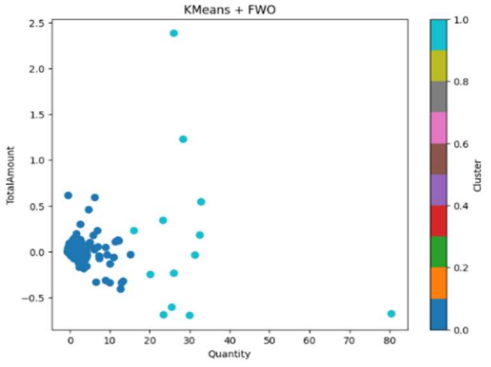


Figure 11. PCA Visualization of K-Means + FWO (OR)

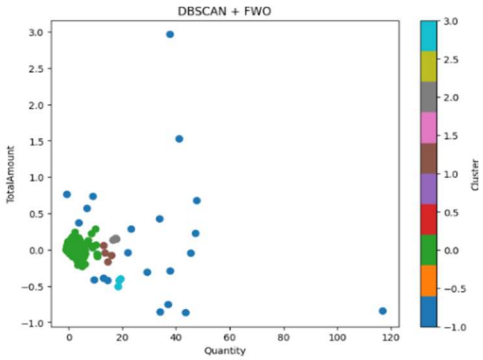


Figure 12. PCA Visualization of DBSCAN + FWO (OR)

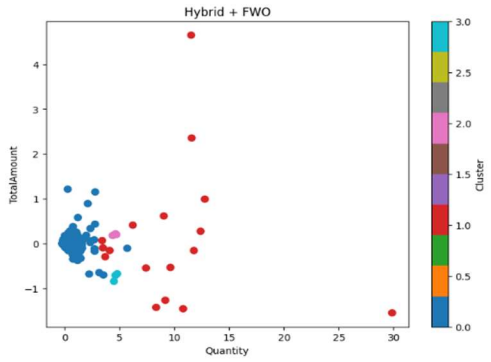


Figure 13. PCA Visualization of Hybrid Method +FWO (OR)

The plot in Figure 11 reveals the most optimal separation pattern and highest cluster compactness corresponding to the K-Means + FWO configuration Although a small number of outliers persist, the overall structure demonstrates improved cohesion and reduced overlap among clusters after feature-weight optimization. This confirms that, for transaction-oriented datasets with pronounced centroid-based structure, a centroid method augmented with adaptive feature weighting yields superior segmentation performance.

3. 3 Discussion

The experimental results demonstrate that applying Feature Weight Optimization (FWO) using a Genetic Algorithm (GA) consistently enhanced clustering quality across both datasets Customer Personality Analysis (CPA) and Online Retail (OR). The Hybrid Clustering + FWO approach achieved the highest Silhouette Score on the CPA dataset (0.9599), while the K-Means + FWO method obtained the best result on the OR dataset (0.9804).

These improvements indicate the adaptive capability of GA in adjusting feature weights proportionally to their importance, ensuring that each attribute contributes optimally to cluster formation. The PCA visualizations corroborate these findings, showing clearer inter-cluster separations and stronger intra-cluster cohesion after optimization.

Compared to previous studies, this approach demonstrates substantial advancement. Kouser et al. [6] confirmed the effectiveness of GA in optimizing feature weighting, while Alsmadi et al. [19] highlighted the superiority of hybrid fuzzy clustering for high-dimensional data. The results of this study extend these findings by applying a hybrid FWO-based model to large-scale consumer segmentation tasks. Additionally, this study aligns with Salman and Fauziah [11], who emphasized K-Means computational efficiency for homogeneous datasets.

However, certain limitations persist notably in the empirical selection of DBSCAN parameters (epsilon and min_samples), which remains dataset-dependent and may influence cluster consistency.

Overall, the findings confirm that the integration of Hybrid Clustering and Feature Weight Optimization is an effective approach to enhance segmentation accuracy in customer analytics. This study reinforces the hypothesis that adaptive feature weighting can improve cluster separation without sacrificing computational efficiency.

The significance of this research lies in the incorporation of evolutionary optimization principles into consumer data segmentation a rarely explored area. Future studies could extend this work by developing deep learning-based adaptive FWO or evolutionary hybrid clustering methods to handle real-time data within distributed computing environments such as Apache Spark or Hadoop. Such developments would further advance clustering intelligence for automated decision-support systems in industrial and e-commerce.

4. CONCLUSION

This study evaluated the performance of K-Means, DBSCAN, and Hybrid Clustering algorithms combined with Feature Weight Optimization (FWO) based on a Genetic Algorithm (GA) to improve the accuracy of customer data segmentation. The experimental results demonstrated that applying FWO significantly enhanced clustering quality, as reflected by higher Silhouette Scores on both datasets Customer Personality Analysis (CPA) and Online Retail (OR). The Hybrid + FWO method achieved the best performance on the CPA dataset with a Silhouette Score of 0.9599, while K-Means + FWO obtained the highest score of 0.9804 on the OR dataset, indicating that data characteristics influence algorithmic effectiveness.

The findings confirm that feature weight optimization using GA can improve both cohesion and separation among clusters, resulting in more adaptive and stable segmentation outcomes. The results also imply that evolutionary optimization provides a viable approach for refining clustering performance in consumer analytics. Future research may explore the integration of deep learning-based FWO and distributed clustering frameworks to enhance efficiency and scalability for large-scale data processing.

REFERENCES

[1] A. A. Rahma, A. Faqih, and A. R. Rinaldi, "Optimalisasi Strategi Pemasaran melalui Segmentasi Pelanggan dengan Analisis RFM dan Algoritma K-Means untuk Bisnis Ritel," *JIKO (Jurnal Informatika dan Komputer)*, vol. 9, no. 2, p. 338, Jun. 2025, doi: 10.26798/jiko.v9i2.1737.

[2] S. D. K. Wardani, A. S. Ariyanto, M. Umroh, and D. Rolliawati, "PERBANDINGAN HASIL METODE CLUSTERING K-MEANS, DB SCANNER & HIERARCHICAL UNTUK ANALISA SEGMENTASI PASAR," *JIKO (Jurnal Informatika dan Komputer)*, vol. 7, no. 2, p. 191, Sep. 2023, doi: 10.26798/jiko.v7i2.796.

[3] Rahmati r and Wijayanto A, "ANALISIS CLUSTER DENGAN ALGORITMA K-MEANS, FUZZY C-MEANS DAN HIERARCHICAL CLUSTERING," *JIKO (Jurnal Informatika dan Komputer)*, vol. 5, no. 2, Mar. 2021.

[4] Z. Wang *et al.*, "AMD-DBSCAN: An Adaptive Multi-density DBSCAN for datasets of extremely variable density," *arXiv preprint arXiv:2210.08162*, 2022, doi: 10.48550/arXiv.2210.08162.

[5] K. N. Sridevi and M. Rajanna, "Hybrid Clustering Framework for Scalable and Robust Query Analysis: Integrating Mini-Batch K-Means with DBSCAN," *International Journal of Advanced Computer Science and Applications*, vol. 16, no. 1, pp. 87–95, 2025, doi: 10.14569/IJACSA.2025.0160187.

[6] K. Kouser, A. Priyam, M. Gupta, S. Kumar, and V. Bhattacharjee, "Genetic Algorithm–Based Optimization of Clustering Algorithms for the Healthy Aging Dataset," *Applied Sciences*, vol. 14, no. 13, p. 5530, 2024, doi: 10.3390/app14135530.

[7] G. Feng, "Feature selection algorithm based on optimized genetic algorithm and the application in high-dimensional data processing," *PLoS One*, vol. 19, no. 5, 2024, doi: 10.1371/journal.pone.0303088.

[8] A. G. Oskouei *et al.*, "Feature-Weighted Fuzzy Clustering Methods: An Experimental Review," *Neurocomputing*, vol. 619, p. 129176, 2025, doi: 10.1016/j.neucom.2024.129176.

[9] M. Gaido, "Distributed Silhouette Algorithm: Evaluating Clustering on Big Data," *arXiv preprint arXiv:2303.14102*, 2023, [Online]. Available: https://arxiv.org/abs/2303.14102

[10] A. Suryaputra Paramita and T. Hariguna, "Comparison of K-Means and DBSCAN Algorithms for Customer Segmentation in E-commerce," *Journal of Digital Marketing and Digital Commerce*, vol. 1, no. 1, pp. 43–62, 2024, doi: 10.47738/jdmdc.v1i1.3.

[11] F. Salman and F. Fauziah, "Comparison Analysis of K-Means and DBSCAN Algorithms for Improving Budget Absorption Efficiency in EIS," *Brilliance: Research of Artificial Intelligence*, vol. 3, no. 2, pp. 378–383, 2023, doi: 10.47709/brilliance.v3i2.3373.

[12] Q.-V. Doan, T. Amagasa, T.-H. Pham, T. Sato, F. Chen, and H. Kusaka, "Structural k-means (Sk-means) and clustering uncertainty evaluation framework (CUEF) for mining climate data," *Geosci Model Dev*, vol. 16, pp. 2215–2233, 2023, doi: 10.5194/gmd-16-2215-2023.

[13] R. Tinós, L. Zhao, F. Chicano, and D. Whitley, "NK Hybrid Genetic Algorithm for Clustering," *arXiv preprint arXiv:2402.03813*, 2024, doi: 10.48550/arXiv.2402.03813.

[14] S. Chowdhury, N. Helian, and R. de Amorim, "Feature weighting in DBSCAN using reverse nearest neighbours," *Pattern Recognit*, vol. 137, p. 109314, 2023, doi: 10.1016/j.patcog.2023.109314.

[15] R. Mussabayev and R. Mussabayev, "Comparative Analysis of Optimization Strategies for K-Means Clustering in Big Data Contexts: A Review," *arXiv preprint arXiv:2310.09819*, 2023, doi: 10.48550/arXiv.2310.09819.

[16] T. Bezdan, Y. Zhang, and Y. Zhang, "Fruit-Fly Algorithm Based Hybrid K-Means Clustering Method for Text Document Clustering," *Mathematics*, vol. 9, no. 16, p. 1929, 2021, doi: 10.3390/math9161929.

[17] P. Bansal *et al.*, "GGA-MLP: A Greedy Genetic Algorithm to Optimize Weights and Biases in MLP," *Contrast Media Mol Imaging*, vol. 2022, p. 4036035, 2022, doi: 10.1155/2022/4036035.

[18] V. V. Baligodugula and F. Amsaad, "Unsupervised Learning: Comparative Analysis of Clustering Techniques on High-Dimensional Data," *arXiv preprint arXiv:2503.23215*, 2025.

[19] M. K. Alsmadi *et al.*, "A Hybrid Topic Modeling Method Based on Dirichlet Multinomial Mixture and Fuzzy Matching Algorithm for Short Text Clustering," *J Big Data*, vol. 11, no. 68, 2024, doi: 10.1186/s40537-024-00930-9.