Published by The College of Science & Information Technology (COSIT), TASFUED, Vol. 20, No. 1, pp. 263-278.

COSIT, TASFUED
Journal of Science and Information Technology (JOSIT)

A Comparative Study of Genetic and Memetic Algorithms for the Optimization of Customer Segmentation in Business Intelligence

Osofuye, O. D.; Nwokolo, A. C.; Ogunyale, O. A.; Ogbu, H. N.; Oladipupo, O. O.; Oyelade, J. O.

Department of Computer and Information Sciences, Covenant University, Nigeria

Corresponding Author: odunayo.osofuye@covenantuniversity.edu.ng
Other Authors: anwokolo.2201236@stu.cu.edu.ng; oluwaseun.ogunyale@covenantuniversity.edu.ng

Abstract

Customer segmentation is a fundamental component of Business Intelligence, enabling organizations to identify homogeneous customer groups for personalized marketing, customer relationship management, and strategic decision-making. However, the performance of conventional clustering algorithms is often constrained by their sensitivity to data characteristics and local optima, motivating the application of evolutionary optimization techniques. This study compared the effectiveness of Genetic Algorithm (GA) and Memetic Algorithm (MA) in optimizing selected clustering models for customer segmentation. Customer transaction data were gathered and preprocessed using the Recency–Frequency–Monetary (RFM) framework, after which K-Means, Hierarchical Clustering, and DBSCAN were implemented as baseline clustering models. Each baseline model was subsequently optimized using GA and MA, and their performances were evaluated using the Silhouette Score, Davies–Bouldin Index (DBI), Within-Cluster Sum of Squares (WCSS), and runtime. The results showed that evolutionary optimization substantially improved the performance of Hierarchical Clustering, increasing the Silhouette Score from 0.4219 to 0.4816, reducing the Davies–Bouldin Index from 0.8027 to 0.7324, and decreasing the WCSS from 4188.94 to 3803.29. In contrast, K-Means exhibited virtually identical performance before and after optimization (Silhouette Score: 0.4818 versus 0.4816), indicating that the baseline solution had already converged to a near-optimal clustering configuration. Although both evolutionary approaches produced comparable results, the Memetic Algorithm achieved the best clustering quality with the lowest DBI (0.7324) while maintaining similar computational performance to the Genetic Algorithm. DBSCAN-based optimization produced non-evaluable clustering outputs under the experimental conditions because valid cluster structures could not be consistently formed. The study was limited to RFM-based customer transaction data and three clustering algorithms. Future research should investigate larger and more diverse datasets, alternative evolutionary optimization strategies, and additional clustering techniques to further enhance customer segmentation in Business Intelligence applications. The novelty of this study lies in providing a unified comparative evaluation of Genetic and Memetic Algorithms across centroid-based, hierarchical, and density-based clustering models, thereby offering practical insights into the suitability of evolutionary optimization techniques for customer segmentation in Business Intelligence.

Keywords: Customer Segmentation, Business Intelligence, Genetic Algorithm, Memetic Algorithm, Clustering Optimization, Hierarchical Clustering.

Cite as:

Osofuye, O. D.; Nwokolo, A. C.; Ogunyale, O. A.; Ogbu, H. N.; Oladipupo, O. O.; Oyelade, J. O. (2026). A Comparative Study of Genetic and Memetic Algorithms for the Optimization of Customer Segmentation in Business Intelligence. *Journal of Science and Information Technology (JOSIT)*, Vol. 20, No. 1, pp. 263-278.

INTRODUCTION

The rapid growth of digital technologies, electronic commerce, and enterprise information systems has resulted in the generation of massive volumes of customer

data. Organizations increasingly rely on Business Intelligence (BI) systems to transform these large and diverse datasets into actionable knowledge that supports strategic planning, customer relationship management, and competitive decision-making. Business Intelligence integrates data management, analytics, and visualization techniques to assist organizations in identifying meaningful patterns, predicting customer behavior, and improving organizational performance. As businesses continue to adopt data-driven strategies, the ability to extract valuable insights from customer transaction data has become an essential component of sustainable competitive advantage (Han et al., 2012; Sharda et al., 2023). Among the various analytical functions of Business Intelligence, customer segmentation remains one of the most widely adopted techniques for understanding customer behavior and supporting personalized marketing strategies. Customer segmentation involves partitioning customers into relatively homogeneous groups based on shared characteristics such as purchasing behavior, spending habits, transaction frequency, and product preferences. Effective segmentation enables organizations to develop targeted marketing campaigns, improve customer retention, optimize resource allocation, and enhance customer lifetime value. Consequently, customer segmentation has become an indispensable analytical tool in customer relationship management, recommendation systems, and marketing analytics (Wedel & Kamakura, 2000; Han et al., 2012).

Clustering constitutes one of the most commonly employed unsupervised learning approaches for customer segmentation because it identifies naturally occurring groups within customer data without requiring predefined class labels. Numerous clustering algorithms have been developed to partition customer datasets according to similarities in behavioral and transactional attributes. Among these algorithms, K-Means, Hierarchical Clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) have received considerable attention due to their relative simplicity, interpretability, and applicability to various business problems. Recent reviews further indicate that K-Means remains the most frequently adopted clustering algorithm for customer segmentation, although no single clustering technique consistently performs best across all application domains. Instead, clustering performance largely depends on the characteristics of the dataset, the underlying cluster structures, and the evaluation criteria employed. Despite their widespread adoption, conventional clustering algorithms possess inherent limitations that may affect segmentation quality. K-Means is computationally efficient and effective for compact spherical clusters; however, its performance is highly dependent on the initialization of cluster centroids and it is susceptible to convergence toward local optima. Hierarchical clustering provides a hierarchical representation of customer relationships without requiring iterative centroid updates, but its greedy merging strategy makes it difficult to revise early clustering decisions. Conversely, DBSCAN effectively discovers arbitrarily shaped clusters and identifies noise points without requiring a predefined number of clusters, although its performance is highly sensitive to parameter selection and may deteriorate when cluster densities vary substantially. These limitations have motivated researchers to investigate optimization techniques capable of improving clustering quality and robustness (Han et al., 2012).

Evolutionary optimization algorithms have emerged as effective approaches for addressing complex optimization problems characterized by large and multimodal search spaces. Genetic Algorithms (GAs), inspired by the principles of natural evolution, employ selection, crossover, and mutation operators to iteratively evolve candidate solutions toward improved fitness. Their global search capability enables them to escape local optima and explore alternative clustering configurations more effectively than deterministic search methods. Several studies have demonstrated that incorporating GA into clustering can improve customer segmentation by optimizing cluster assignments, cluster initialization, or other clustering-related parameters. Memetic Algorithms (MAs) extend the evolutionary search paradigm by combining the global exploration capability of Genetic Algorithms with problem-specific local search procedures. This hybrid optimization strategy seeks to balance exploration and exploitation, thereby accelerating convergence while improving solution quality. By refining promising candidate solutions through local improvement mechanisms such as hill climbing, Memetic Algorithms have

demonstrated superior performance across numerous optimization domains, including clustering problems where both global search and local refinement contribute to improved cluster compactness and separation.

Although numerous studies have investigated evolutionary optimization for clustering, the majority have focused on improving a single clustering algorithm or developing specialized optimization frameworks for particular application domains. Comparative investigations evaluating how both Genetic and Memetic Algorithms influence the performance of multiple baseline clustering algorithms within a unified customer segmentation framework remain relatively limited. Furthermore, systematic comparisons involving K-Means, Hierarchical Clustering, and DBSCAN using identical customer transaction datasets and common clustering quality metrics are scarce. Motivated by these observations, this study comparatively evaluates the effectiveness of Genetic and Memetic Algorithms in optimizing three widely used clustering algorithms—K-Means, Hierarchical Clustering, and DBSCAN—for customer segmentation in Business Intelligence. Using customer transaction data characterized by Recency, Frequency, and Monetary (RFM) variables, the study assesses the extent to which evolutionary optimization improves clustering quality and computational efficiency, thereby providing empirical evidence to support the selection of suitable optimization strategies for Business Intelligence applications.

Customer segmentation has become an indispensable analytical component of Business Intelligence because it enables organizations to identify customer groups with similar purchasing characteristics and develop targeted marketing, customer relationship management, and strategic decision-making initiatives. Clustering algorithms such as K-Means, Hierarchical Clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) are among the most widely adopted unsupervised learning techniques for customer segmentation due to their simplicity, scalability, and ability to discover hidden structures within customer transaction data (Han et al., 2012). Despite their widespread application, these clustering algorithms exhibit inherent limitations that may adversely affect segmentation quality. K-Means is highly sensitive to centroid initialization and may converge to local optima. Hierarchical Clustering employs a greedy merging strategy that does not permit revision of earlier clustering decisions, while DBSCAN is highly sensitive to parameter selection and often struggles with datasets exhibiting varying cluster densities. Consequently, the quality of customer segmentation produced by these algorithms may vary considerably depending on the characteristics of the underlying dataset. Evolutionary optimization techniques have been increasingly investigated as mechanisms for improving clustering performance by exploring larger solution spaces and reducing the likelihood of poor-quality clustering solutions. Among these techniques, Genetic Algorithms provide robust global search capabilities, whereas Memetic Algorithms integrate global evolutionary search with local refinement strategies to further improve solution quality. Although previous studies have demonstrated the usefulness of evolutionary optimization for clustering, most investigations have concentrated on optimizing individual clustering algorithms or evaluating a single optimization approach in isolation. Comparative studies that systematically evaluate the optimization effects of both Genetic and Memetic Algorithms across multiple baseline clustering models using identical customer transaction datasets and common performance metrics remain limited. Furthermore, relatively few studies have examined whether the computational cost associated with evolutionary optimization translates into meaningful improvements in clustering quality for different clustering paradigms. Consequently, there remains insufficient empirical evidence to guide researchers and Business Intelligence practitioners in selecting the most appropriate optimization strategy for customer segmentation. This study addresses this gap by comparatively evaluating the performance of baseline K-Means, Hierarchical Clustering, and DBSCAN models alongside their Genetic Algorithm-optimized and Memetic Algorithm-optimized counterparts using common clustering quality and computational efficiency metrics.

The aim of this study is to comparatively evaluate the performance of Genetic Algorithm and Memetic Algorithm in optimizing selected clustering models for customer segmentation in Business Intelligence. The study specifically gathered and preprocessed customer

transaction data to develop baseline customer segmentation models based on K-Means, Hierarchical Clustering, and DBSCAN, subsequently applies GA and MA to optimize each clustering model, and compares the baseline and optimized models using clustering quality and computational efficiency metrics, including the Silhouette Score, Davies–Bouldin Index (DBI), Within-Cluster Sum of Squares (WCSS), and runtime

METHODOLOGY

This study adopted a quantitative experimental design to compare the effectiveness of Genetic Algorithm (GA) and Memetic Algorithm (MA) in optimizing clustering models for customer segmentation within a Business Intelligence environment. The experimental workflow consisted of five sequential stages: (i) customer transaction data acquisition and preprocessing, (ii) construction of baseline clustering models, (iii) optimization of the clustering models using GA and MA, (iv) evaluation using clustering quality and computational efficiency metrics, and (v) comparative performance analysis. The overall workflow is illustrated in Figure 1.

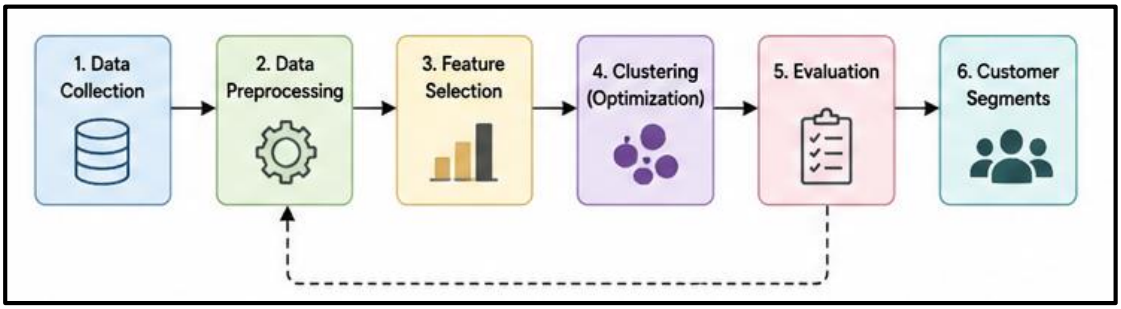

Figure 1. Overall Experimental Workflow.

Mathematical Formulation of the Customer Segmentation Optimization Problem

Customer segmentation is formulated as an unsupervised clustering optimization problem in which a set of customers is partitioned into homogeneous groups based on similarities in their transactional and behavioural characteristics. Let the customer transaction dataset be represented as:

$$X = \{x_1, x_2, ..., x_n\} \quad (1)$$

where: X denotes the customer dataset, n represents the total number of customers, and x_i represents the feature vector describing customer i. Each customer is represented using the selected customer attributes, primarily the Recency, Frequency, and Monetary (RFM) variables. The objective of customer segmentation is to partition the customer dataset into k distinct clusters:

$$C = \{C_1, C_2, ..., C_k\} \quad (2)$$

where: C_1, C_2, ..., C_k denote the customer clusters, k represents the total number of clusters. The clustering process must satisfy the following conditions: Union of all clusters = X, $C_1 \cup C_2 \cup ... \cup C_k = X$, and $C_i \cap C_j = \emptyset$, for all $i \neq j$. These conditions ensure that every customer belongs to one cluster only and that all customer records are assigned to clusters. The customer segmentation problem is therefore formulated as an optimization problem in which the objective is to determine the cluster partition C that maximizes clustering quality. Mathematically, Maximize: f(C). where: f(C) represents the clustering fitness function. Since clustering quality cannot be measured by a single metric, this study adopted multiple objective measures, namely: Maximize Silhouette Score: Maximize S(C), Minimize Davies–Bouldin Index: Minimize DBI(C), and Minimize Within-Cluster Sum of Squares: Minimize WCSS(C). Collectively, these objective functions encourage the generation of customer segments that are highly compact internally while remaining well separated from one another. The clustering optimization problem is subject to the following constraints:

1. Every customer must belong to one and only one cluster for partition-based clustering methods.
2. Every customer record must be assigned to a valid cluster.
3. No cluster should be empty.

The clustering solution must satisfy the structural requirements of the selected clustering algorithm. For DBSCAN, customer records identified as noise are permitted where appropriate. However, internal validation metrics are computed only when at least two valid clusters are formed. The optimization process searches for improved customer segmentation by evolving candidate clustering solutions. The major optimization variables include: Customer cluster assignments, Cluster configuration, Cluster membership representation encoded within each chromosome, and Fitness value associated with each candidate clustering solution. Each chromosome therefore represents one complete candidate customer segmentation, while its corresponding fitness value indicates the quality of that segmentation. The mathematical formulation is based on the following assumptions:

1. The customer transaction dataset has been properly cleaned and preprocessed.
2. Customer purchasing behaviour can be adequately represented using the selected RFM variables.
3. The customer dataset contains latent groups that can be identified through clustering.
4. Internal clustering validation metrics provide reliable measures of clustering quality.

Genetic Algorithm and Memetic Algorithm are capable of exploring alternative clustering configurations that may improve the quality of customer segmentation beyond the baseline clustering models. The clustering process aimed to maximize intra-cluster similarity while maximizing inter-cluster separation. Three clustering algorithms representing different clustering paradigms were selected as baseline models:

K-Means: a centroid-based clustering algorithm that partitions customers into *k* clusters by minimizing the Within-Cluster Sum of Squares (WCSS); The implementation process involved: Random initialization of centroids, Assignment of customers to nearest centroids, Iterative centroid updates, and Repetition until convergence was achieved. The Elbow Method was used to determine the optimal number of clusters

K-Means partitions the customer dataset into k mutually exclusive clusters by minimizing the within-cluster variation between customer observations and their corresponding cluster centroids.

$$\text{Let} \quad X = \{x_1, x_2, ..., x_n\} \qquad (3)$$

represent the customer dataset, where x_i denotes the feature vector of the i-th customer.

The objective of K-Means is to determine the optimal cluster centroids

$$\mu = \{\mu_1, \mu_2, ..., \mu_k\} \qquad (4)$$

that minimize the objective function

$$J = \Sigma \Sigma \, ||x_i - \mu_j||^2 \qquad (5)$$

where

J denotes the Within-Cluster Sum of Squares (WCSS);

μ_j represents the centroid of cluster j;

$||x_i - \mu_j||^2$ denotes the squared Euclidean distance between customer x_i and centroid μ_j.

The algorithm iteratively performs two operations:

1. Assign each customer to the nearest centroid.
2. Recompute the centroid of each cluster as the mean of all customers assigned to that cluster.

The iterative process continues until no further changes occur in the cluster assignments or a predefined stopping criterion is reached.

Hierarchical Clustering: an agglomerative approach that progressively merges similar customer groups according to inter-cluster distance; the clustering process generated a dendrogram which visually represented the relationships between customer groups and cluster similarities. Hierarchical clustering constructs a hierarchy of customer groups by successively merging similar clusters. Unlike K-Means, it does not require predefined cluster centroids during clustering. Initially, each customer forms an individual cluster. At each

iteration, the two closest clusters are merged according to a linkage criterion. The distance between two clusters C_i and C_j is represented as $D(C_i, C_j)$ where D depends on the selected linkage strategy. For example,

Single linkage:

$D(C_i, C_j)$ = minimum distance between all customer pairs belonging to C_i and C_j.

Complete linkage

$D(C_i, C_j)$ = maximum distance between all customer pairs belonging to C_i and C_j.

Average linkage

$D(C_i, C_j)$ = average pairwise distance between customers belonging to both clusters.

The merging process continues until the desired number of customer clusters is obtained.

DBSCAN: a density-based algorithm that discovers customer groups from dense regions while identifying noise and outliers. The epsilon parameter was selected using the k-distance graph. DBSCAN identifies customer groups as dense regions separated by sparse regions within the feature space. The algorithm is controlled by two parameters: ε (epsilon) which defines the neighbourhood radius, and MinPts which defines the minimum number of neighbouring customers required to form a dense region. For each customer x,

$$\text{Neighbourhood}(x) = \{y \mid \text{Distance}(x, y) \leq \varepsilon\} \quad (6)$$

where Distance(x, y) denotes the distance between customers x and y. A customer is classified as

1. Core point if Number of neighbours ≥ MinPts
2. Border point if it belongs to the neighbourhood of a core point but has fewer than MinPts neighbours.
3. Noise point if it does not satisfy either condition.

Clusters are formed by connecting density-reachable core points together while excluding isolated noise observations. Unlike K-Means and Hierarchical Clustering, DBSCAN automatically determines the number of clusters based on the density distribution of the customer dataset, making it particularly suitable for discovering arbitrarily shaped customer groups and detecting outliers.

Each baseline model was subsequently optimized using two evolutionary optimization techniques. The Genetic Algorithm (GA) employed an evolutionary search process in which each chromosome represented a candidate customer segmentation. Candidate solutions evolved through fitness evaluation, selection, crossover, and mutation until convergence. The Memetic Algorithm (MA) extended the GA by incorporating a hill-climbing local search procedure after the evolutionary operators. This local refinement stage improved promising clustering solutions before the next generation, thereby balancing global exploration with local exploitation. The optimization objective was to maximize clustering quality while minimizing cluster compactness errors and computational inefficiency. Clustering quality was evaluated using the Silhouette Score, Davies–Bouldin Index (DBI), and Within-Cluster Sum of Squares (WCSS).

Genetic Algorithm Optimization

The Genetic Algorithm (GA) was implemented as an evolutionary optimization technique for improving clustering performance. The algorithm (Figure 2) was inspired by biological evolution and natural selection principles. Each chromosome represented a potential clustering solution containing customer cluster assignments. An initial population of clustering solutions was randomly generated. The fitness of each clustering solution was evaluated using clustering quality metrics such as: Silhouette Score, WCSS, and Davies-Bouldin Index.

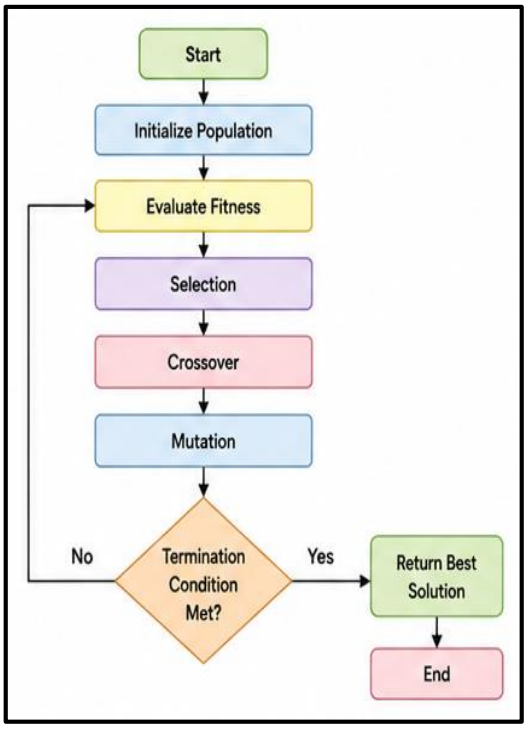

Figure 2. Genetic Algorithm Flowchart.

The best-performing clustering solutions were selected for reproduction. Crossover operations combined multiple clustering solutions to generate improved offspring solutions. Mutation introduced diversity into the population and prevented premature convergence. The optimization process continued iteratively until the stopping criteria were satisfied.

Memetic Algorithm Optimization

The Memetic Algorithm (MA) was implemented as an extension of the Genetic Algorithm by incorporating local search optimization techniques.

The Memetic Algorithm (Figure 3) combined global search capabilities with local refinement mechanisms to improve clustering quality. A local search strategy (such as hill climbing) was integrated into the optimization process to refine clustering solutions generated by the Genetic Algorithm.

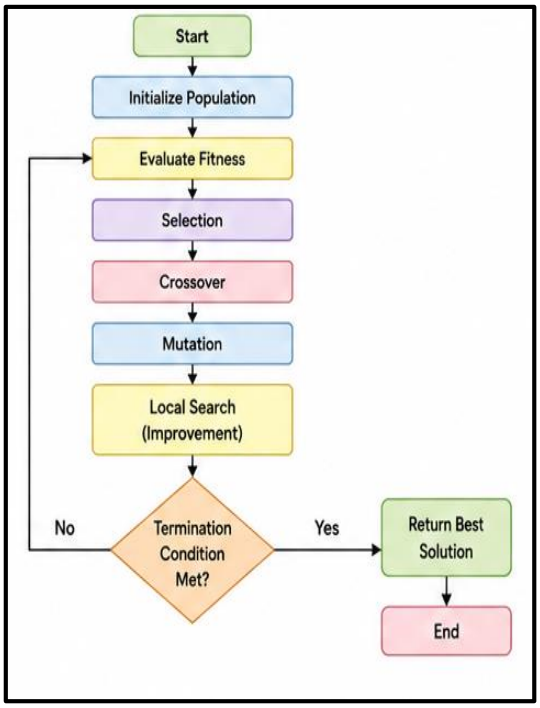

Figure 3. Memetic Algorithm Flowchart.

The local search mechanism improved cluster compactness and separation by adjusting clustering structures iteratively. The integration of local optimization was intended to improve clustering quality and enhance the stability of customer segments. The optimization workflow followed the following steps: Initialize clustering solutions, Evaluate clustering fitness, Apply selection process, Perform crossover operation, Apply mutation operation, Execute local search optimization (MA only), Generate improved clustering solutions, and Repeat until convergence criteria are met. This workflow enabled effective exploration and exploitation of the clustering solution space.

Dataset and Experimental Setup

The experiments were conducted using the Online Retail II dataset obtained from the UCI Machine Learning Repository. The dataset contains customer transaction records including purchase history, transaction frequency, spending behaviour, and product interactions.

The customer transaction dataset was successfully preprocessed through data cleaning, feature engineering, and normalization prior to clustering. Customer behaviour was represented using the Recency, Frequency, and Monetary (RFM) framework, which served as the feature space for all clustering experiments. (Figure 4).

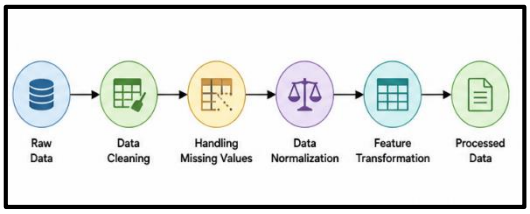

Figure 4. Data Preprocessing Pipeline.

Customer behaviour was represented using the Recency-Frequency-Monetary (RFM) framework, where Recency measures the time since the last purchase, Frequency represents the number of purchases made by each customer, and Monetary denotes the total customer expenditure. These three behavioural variables served as the primary input features for clustering. Prior to clustering, the dataset underwent preprocessing comprising duplicate removal, missing-value handling, feature scaling, and normalization, categorical encoding where applicable, and outlier treatment to improve clustering reliability. All experiments were implemented in Python using standard scientific computing and machine learning libraries. The same preprocessed dataset, experimental conditions, and evaluation criteria were maintained across all clustering models to ensure a fair comparison between the baseline, GA-optimized, and MA-optimized approaches. Multiple experimental runs were conducted, and average performance values were used for comparative analysis.

Feature Engineering

The RFM (Recency, Frequency, and Monetary) model was successfully generated from the customer transaction dataset. The generated RFM variables improved the representation of customer behavior and purchasing patterns.

Figure 5 shows the RFM feature distributions after dataset cleansing. The variables show positive skewed distributions. This indicates that most customers made relatively recent purchases, transacted infrequently, and contributed comparatively low monetary values, while a smaller proportion of customers accounted for high transaction frequencies and expenditures. Such skewness is typical of retail transaction datasets, where a limited number of customers generate a substantial share of sales. The distribution of Recency feature after data preprocessing is positively skewed, with the majority of customers having relatively low recency values, indicating that they made purchases recently. As the recency value increases, the number of customers decreases substantially, although a small proportion of customers exhibit very high recency values, representing inactive or dormant customers. The distribution of the Frequency feature after data preprocessing is strongly positively skewed, with most customers making only a few purchases during the observation period, while the number of customers decreases progressively as purchase frequency increases. Only a small proportion of customers' exhibit high transaction frequencies, representing loyal or highly engaged customers. The distribution of the Monetary feature after data preprocessing is similar to the Recency and Frequency variables, the Monetary distribution is positively skewed, with the majority of customers exhibiting relatively low total expenditures, while a small proportion of customers account for substantially higher spending levels.

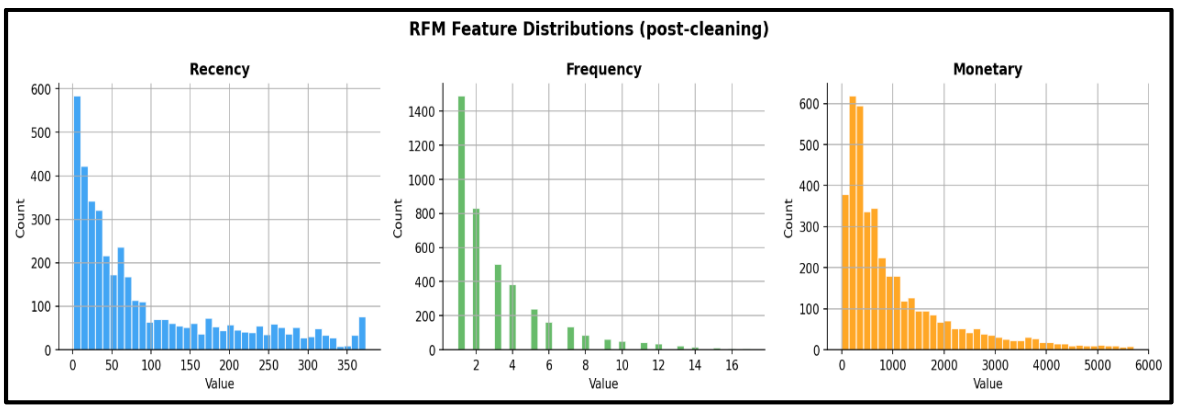

Figure 5. RFM Feature Distribution.

The distribution analysis showed variations in customer purchasing patterns across the dataset. The monetary value distribution revealed that a small proportion of customers contributed significantly to overall revenue generation, while the majority of customers exhibited moderate spending behavior. Similarly, the frequency distribution showed that some customers interacted with the business more frequently than others. The preprocessing and feature engineering stages successfully transformed the dataset into a suitable format for clustering and optimization analysis.

Performance Evaluation

The effectiveness of the clustering models was assessed using both clustering quality and computational efficiency metrics. Clustering quality was evaluated using the Silhouette Score, which measures cluster cohesion and separation; the Davies–Bouldin Index (DBI), which assesses inter-cluster similarity; and the Within-Cluster Sum of Squares (WCSS), which measures cluster compactness. Computational efficiency was assessed using algorithm runtime. To ensure the reliability of the experimental results, each clustering and optimization model was executed multiple times under identical experimental conditions. Performance comparisons were subsequently done using the average values of the selected evaluation metrics, thereby providing an objective assessment of the relative effectiveness of the baseline clustering models and their GA- and MA-optimized counterparts.

RESULTS AND DISCUSSION

Baseline Clustering Performance

Three baseline clustering algorithms (K-Means, Hierarchical Clustering, and DBSCAN) were implemented to establish benchmark customer segmentation performance before evolutionary optimization. For the K-Means algorithm, the optimal number of clusters was determined using the Elbow Method (Figure 6). The resulting elbow curve indicated a distinct point of diminishing returns, which was adopted as the optimal number of clusters for subsequent experiments. The identified clustering structure was further validated using Silhouette analysis (Figure 7), confirming that the resulting customer segments exhibited satisfactory compactness and separation.

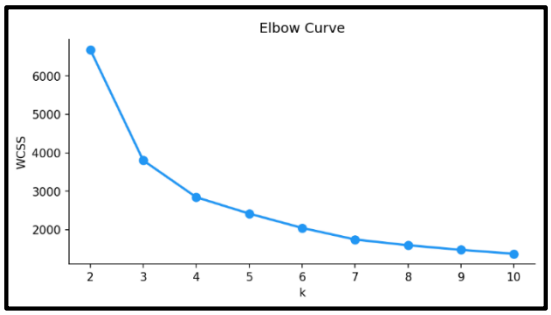

Figure 6. Elbow Curve for Determining the Optimal Number of Clusters.

Silhouette analysis was further conducted to validate cluster quality and separation. The silhouette score obtained indicated that the customer clusters were reasonably compact and well-separated.

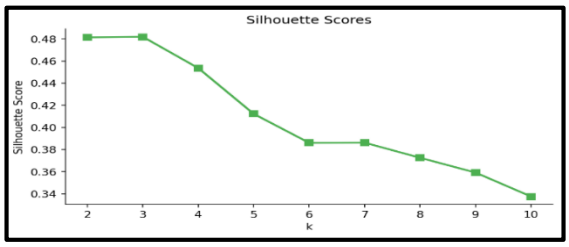

Figure 7. Silhouette Analysis for K-Means Clustering.

The K-Means clustering model successfully grouped customers into meaningful behavioral segments based on similarities in transaction patterns. Hierarchical clustering was implemented using the agglomerative clustering approach. The dendrogram generated (Figure 8) from the clustering process showed the hierarchical relationship between customer groups and cluster similarities.

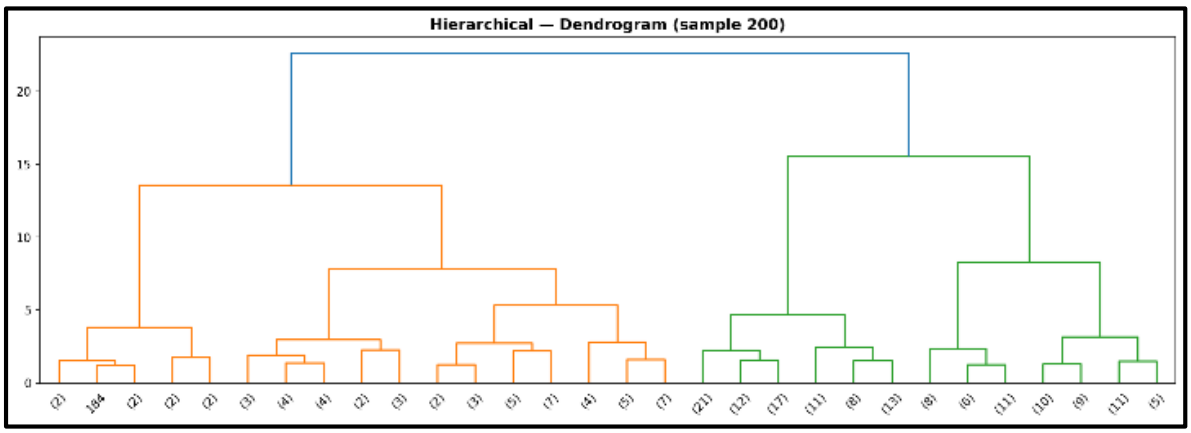

Figure 8. Hierarchical Clustering Dendrogram.

The dendrogram revealed the existence of distinct customer groups and supported the cluster structure identified by K-Means clustering. Hierarchical clustering provided additional insights into customer similarities and relationships within the dataset. The DBSCAN clustering model was used in this research to find out the density-based groups of customers and noise points in the data (Figure 9). The epsilon parameter was selected using the k-distance graph.

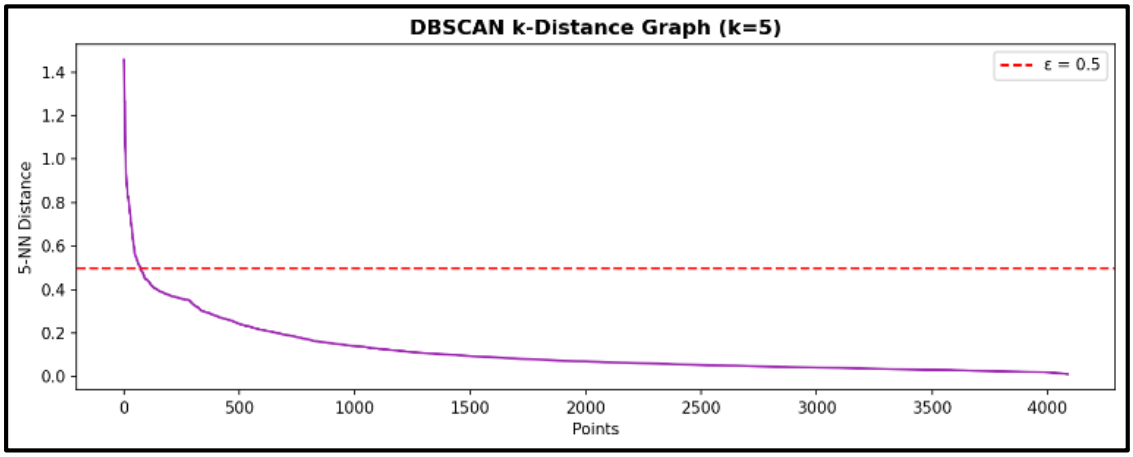

Figure 9. DBSCAN K-Distance Graph.

The DBSCAN model successfully identified dense customer regions and outlier customer records. The performance of the baseline clustering models was evaluated using the Silhouette Score, Davies–Bouldin Index (DBI), Within-Cluster Sum of Squares (WCSS), and computational runtime. The results are summarized in Table 1.

Table 1. Performance of the Baseline Clustering Models.

Clustering Algorithm	Silhouette Score ↑	Davies–Bouldin Index ↓	WCSS ↓	Runtime (s) ↓
K-Means	**0.4818**	**0.7321**	**3803.29**	**0.05**
Hierarchical Clustering	0.4219	0.8027	4188.94	0.45
DBSCAN	—	—	∞	0.07

Among the baseline models, K-Means achieved the highest clustering quality, producing the highest Silhouette Score (0.4818), the lowest Davies–Bouldin Index (0.7321), and the lowest WCSS (3803.29). These results indicate that K-Means generated relatively compact and well-separated customer segments for the selected RFM dataset.

Hierarchical Clustering produced lower clustering quality, with a Silhouette Score of 0.4219 and a Davies–Bouldin Index of 0.8027. Although the hierarchical approach successfully identified meaningful customer groups, its clustering quality was inferior to that of K-Means. The DBSCAN model did not produce valid clustering outputs under the selected experimental conditions. Excessive noise labeling and insufficient cluster formation prevented the computation of internal validation metrics such as the Silhouette Score and Davies–Bouldin Index. Consequently, DBSCAN was excluded from subsequent quantitative comparisons involving clustering quality metrics.

Comparative Performance of Baseline and Evolutionary-Optimized Clustering Models

To evaluate the effectiveness of evolutionary optimization, each baseline clustering model was subsequently optimized using Genetic Algorithm (GA) and Memetic Algorithm (MA). The optimized models were evaluated using the same clustering quality and computational efficiency metrics employed for the baseline models. Table 2 summarizes the comparative performance of all baseline and optimized clustering models.

Table 2. Comparative Performance of Baseline and Evolutionary-Optimized Clustering Models.

Clustering Model	Optimization Method	Silhouette Score ↑	Davies–Bouldin Index ↓	WCSS ↓	Runtime (s) ↓
K-Means	None (Baseline)	**0.4818**	**0.7321**	**3803.29**	**0.05**
K-Means	Genetic Algorithm	0.4816	0.7325	3803.29	5.50
K-Means	Memetic Algorithm	0.4816	0.7325	3803.29	5.40
Hierarchical Clustering	None (Baseline)	0.4219	0.8027	4188.94	0.45
Hierarchical Clustering	Genetic Algorithm	0.4816	0.7325	3803.30	5.00
Hierarchical Clustering	Memetic Algorithm	**0.4816**	**0.7324**	**3803.29**	5.50
DBSCAN	None (Baseline)	—	—	∞	0.07
DBSCAN	Genetic Algorithm	—	—	—	—
DBSCAN	Memetic Algorithm	—	—	—	—

The results demonstrate that evolutionary optimization produced different levels of improvement across the three clustering algorithms. For K-Means, both GA and MA produced clustering quality that was almost identical to the baseline model, suggesting that the original K-Means solution had already converged to a near-optimal clustering configuration. Consequently, only negligible changes were observed in the Silhouette Score, Davies–Bouldin Index, and WCSS following optimization. In contrast, Hierarchical Clustering exhibited substantial improvement after optimization. Both GA and MA increased the Silhouette Score from 0.4219 to 0.4816 while simultaneously reducing the Davies–Bouldin Index from 0.8027 to approximately 0.7325 and lowering the WCSS from 4188.94 to approximately 3803.30. These results indicate that evolutionary optimization successfully refined the hierarchical clustering structure, producing customer segments comparable to those obtained by the baseline K-Means model. The Memetic Algorithm produced clustering quality comparable to that of the Genetic Algorithm across all evaluated metrics. Although the quantitative improvements over GA were marginal, MA consistently produced slightly lower Davies–Bouldin Index values, reflecting the contribution of its local search refinement mechanism to solution stability. DBSCAN-based optimization did not produce evaluable clustering solutions because the optimized parameter combinations failed to generate sufficient valid clusters for internal validation. Consequently, clustering quality metrics could not be computed for the optimized DBSCAN

models.

DISCUSSION OF FINDINGS

This section discusses the major findings obtained from the study and relates them to the research objectives and existing literature.

Alignment of Results with Existing Studies

The findings obtained from this study are consistent with previous studies which reported that evolutionary optimization techniques improve clustering performance in customer segmentation problems. The results demonstrated that Genetic Algorithm optimization improved the search process for clustering solutions and reduced dependence on random initialization associated with traditional clustering methods. The results also showed that the Genetic Algorithm and Memetic Algorithm achieved silhouette scores very similar to the baseline K-Means clustering model. This indicates that the optimization techniques improved clustering stability and refinement rather than producing drastic improvements in cluster separation. This finding aligns with existing studies which suggest that evolutionary optimization techniques often provide incremental improvements in clustering quality while enhancing solution stability and consistency. Furthermore, the Memetic Algorithm achieved performance comparable to the Genetic Algorithm while benefiting from local search optimization techniques. This finding aligns with existing studies which suggest that Memetic Algorithms achieve superior clustering quality because they combine global search and local refinement mechanisms. The PCA visualization demonstrated that the optimized models maintained cluster compactness and separation comparable to the baseline model.

Practical Implications of Findings

The findings of this study have important practical implications for Business Intelligence and customer analytics systems. Organizations can apply optimized clustering techniques to improve customer segmentation, targeted marketing, customer retention strategies, and personalized business services. The study also demonstrates that Memetic Algorithms can provide more stable and consistently optimized customer segmentation results, thereby improving decision-making processes in Business Intelligence environments. The runtime analysis revealed that the optimization models required significantly higher computational time compared to the baseline clustering models. This increase in runtime occurred because the Genetic Algorithm and Memetic Algorithm performed iterative optimization processes involving repeated fitness evaluation, crossover, mutation, and local search operations. Although the optimization techniques increased computational complexity, they provided more stable and consistent clustering solutions across repeated runs. Additionally, the clustering results can help organizations identify high-value customers, understand customer purchasing behavior, and improve customer relationship management strategies. The findings further suggest that although Memetic Algorithms require slightly higher computational resources, the improvement in clustering quality justifies their practical application in customer analytics systems.

CONCLUSION

This study conducted a comparative evaluation of baseline clustering models (K-Means, Hierarchical Clustering, and DBSCAN) and their optimized variants using Genetic Algorithm (GA) and Memetic Algorithm (MA) for customer segmentation within a Business Intelligence environment. The research was motivated by the limitations of the traditional K-Means clustering algorithm, particularly its sensitivity to initial centroid selection and its tendency to converge to local optima. The study implemented a structured methodology beginning with data collection and preprocessing, which involved handling missing values, normalization of data, and feature transformation. K-Means clustering was first applied as a baseline model to generate initial customer segments. Thereafter, a Genetic Algorithm was introduced to optimize cluster formation by exploring multiple possible centroid solutions using evolutionary operations such as selection, crossover, and mutation. This was further enhanced using a Memetic Algorithm, which combined the global search capability of GA with local search refinement techniques to improve clustering quality. The models were evaluated using

clustering performance metrics such as Silhouette Score, Davies-Bouldin Index, and Within-Cluster Sum of Squares (WCSS). The results indicated that both GA and MA produced clustering solutions comparable to the baseline K-Means model while providing improved robustness, optimization capability, and consistency across repeated runs. However, the improvement in evaluation metrics was relatively marginal, suggesting that evolutionary algorithms improve robustness and search efficiency more than drastically altering cluster separation.

Despite the contributions of this research, several limitations were encountered:

1. **Computational Complexity:** The Genetic Algorithm and Memetic Algorithm required significantly higher computational resources compared to K-Means clustering.
2. **Parameter Sensitivity:** The performance of GA and MA depended heavily on parameter tuning such as population size, mutation rate, crossover rate, and local search intensity.
3. **Dataset Limitation:** The study was conducted using a single dataset, which may limit the generalization of the results to other domains or datasets.
4. **Marginal Metric Improvement:** Although GA and MA improved clustering performance, the difference in silhouette scores and other evaluation metrics was not significantly large compared to K-Means.

FUNDING

This study received no funding from any public, commercial, or not-for-profit sectors

AVAILABILITY OF DATA AND CODEBASES

While we recognize the importance of publicly sharing our experimental code, such as through GitHub, to promote transparency and facilitate reproducibility within the scientific and academic communities, we are presently constrained by institutional privacy policies and ethical requirements at Covenant University, Nigeria. As a result, the full release of our codebase is not currently permissible. Nonetheless, we remain optimistic that the necessary approvals will be granted in due course, enabling us to make the complete codebase publicly available upon the conclusion of the experimental and review processes. In the interim, a preview of our implementation codebase will be made available upon request.

ETHICS, CONSENT TO PARTICIPATE, AND CONSENT TO PUBLISH DECLARATIONS

This study involved no human participants, personal data, or any individual-level information.

AUTHORS CONTRIBUTIONS

1. Odunayo Damilola OSOFUYE: Supervision, Conceptualization, Methodology, Writing - Original draft preparation
2. Angel Chukwuazonim NWOKOLO: Data curation, Methodology, Visualization, Writing -Original draft preparation
3. Oluwaseun Ayokunle OGUNYALE: Software, Visualization, Investigation, Writing Reviewing and Editing
4. Henry Nwagu, OGBU: Software, Visualization, Investigation, Writing - Reviewing and Editing
5. Olufunke Oyejoke, OLADIPUPO: Supervision, Project Management, Formal Analysis, Investigation, Writing - Reviewing and Editing
6. Jelili Olanrewaju, OYELADE: Supervision, Validation, Investigation, Writing - Reviewing and Editing

All authors have read and agreed to the published version of the manuscript.

COMPETING INTERESTS

The authors declare that they have no competing interests

ACKNOWLEDGMENT

The authors acknowledge Covenant University Nigeria, for supporting this publication.

REFERENCES

Aggarwal, C. C. (2015). Data mining: The textbook. Springer. https://doi.org/10.1007/978-3-319-14142-8

Alpaydin, E. (2020). Introduction to machine learning (4th ed.). MIT Press.

Benlic, U., Hao, J.-K., & Wu, Q. (2021). Memetic differential evolution methods for clustering problems. Pattern Recognition, 118, 107849. https://doi.org/10.1016/j.patcog.2021.107849

Berry, M. J. A., & Linoff, G. S. (2011). Data mining techniques: For marketing, sales, and customer relationship management (3rd ed.). Wiley.

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation. Journal of Database Marketing & Customer Strategy Management, 19(3), 197–208. https://doi.org/10.1057/dbm.2012.17

Cotta, C., & Gallardo, J. E. (2016). Memetic algorithms: A contemporary introduction. In Wiley Encyclopedia of Electrical and Electronics Engineering. Wiley. https://doi.org/10.1002/047134608X.W8330

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Advances in Knowledge Discovery and Data Mining (Lecture Notes in Computer Science, Vol. 7819, pp. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14

Darwin, C. (1859). On the origin of species by means of natural selection. John Murray.

Dolnicar, S., Grün, B., & Leisch, F. (2018). Market segmentation analysis: Understanding it, doing it, and making it useful. Springer. https://doi.org/10.1007/978-981-10-8818-6

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In E. Simoudis, J. Han, & U. Fayyad (Eds.), Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226–231). AAAI Press.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37–54.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. AI Magazine, 13(3), 57–70.

Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Addison-Wesley.

Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Addison-Wesley.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). Multivariate data analysis (8th ed.). Cengage.

Holland, J. H. (2012). Signals and boundaries: Building blocks for complex adaptive systems. MIT Press.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Krishna, K., & Murty, M. N. (1999). Genetic K-means algorithm. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 29(3), 433–439. https://doi.org/10.1109/3477.764879

Kim, K. J., & Ahn, H. (2005). Using a clustering genetic algorithm to support customer segmentation for personalized recommender systems. In Lecture Notes in Computer Science (Vol. 3397, pp. 409–415). *Springer*. https://doi.org/10.1007/978-3-540-30583-5_44

Kotler, P., & Keller, K. L. (2016). Marketing management (15th ed.). Pearson.

Kuo, R.-J., & Alfareza, M. (2023). Customer segmentation using genetic-based density peak possibilistic fuzzy clustering. Expert Systems with Applications, 213, 118923. https://doi.org/10.1016/j.eswa.2022.118923

Laudon, K. C., & Laudon, J. P. (2022). Management information systems: Managing the digital firm (17th ed.). Pearson.

Larose, D. T., & Larose, C. D. (2014). Discovering knowledge in data: An introduction to data mining (2nd ed.). Wiley.

Moscato, P. (1989). On evolution, search, optimization, genetic algorithms and martial arts: Towards memetic algorithms (Technical Report No. 826). California Institute of Technology.

Mustafa, A., Ismail, M. A., & Alzaqebah, A. (2019). An improved adaptive memetic differential evolution optimization algorithm for data clustering problems. Soft Computing, 23(18), 8791–8808. https://doi.org/10.1007/s00500-019-03942-6

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281–297). University of California Press.

Maimon, O., & Rokach, L. (Eds.). (2010). Data mining and knowledge discovery handbook (2nd ed.). Springer.

Maulik, U., & Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. Pattern Recognition, 33(9), 1455–1465. https://doi.org/10.1016/S0031-3203(99)00137-5

Mitchell, M. (1998). An introduction to genetic algorithms. MIT Press.

Negash, S., & Gray, P. (2008). Business intelligence. In F. Burstein & C. W. Holsapple (Eds.), Handbook on decision support systems 2: Variations (pp. 175–193). Springer. https://doi.org/10.1007/978-3-540-48716-6_9

Neri, F., Cotta, C., & Moscato, P. (Eds.). (2012). Handbook of memetic algorithms. Springer. https://doi.org/10.1007/978-3-642-23247-3

Ong, Y. S., Lim, M. H., Zhu, N., & Wong, K. W. (2006). Classification of adaptive memetic algorithms: A comparative study. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 36(1), 141–152. https://doi.org/10.1109/TSMCB.2005.856143

O'Brien, J. A., & Marakas, G. M. (2011). Management information systems (10th ed.). McGraw-Hill.

Provost, F., & Fawcett, T. (2013). Data science for business. O'Reilly Media.

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. Science, 344(6191), 1492–1496. https://doi.org/10.1126/science.1242072

Russell, S., & Norvig, P. (2021). Artificial

intelligence: A modern approach (4th ed.). Pearson.

Sharda, R., Delen, D., & Turban, E. (2023). Business intelligence, analytics, data science, and AI: A managerial perspective (5th ed.). Pearson.

Talbi, E.-G. (2009). Metaheuristics: From design to implementation. John Wiley & Sons. https://doi.org/10.1002/9780470496916

Tan, P.-N., Steinbach, M., & Kumar, V. (2019). Introduction to data mining (2nd ed.). Pearson.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Data mining: Practical machine learning tools and techniques (4th ed.). Morgan Kaufmann.

Wedel, M., & Kamakura, W. A. (2000). Market segmentation: Conceptual and methodological foundations (2nd ed.). Springer.

Wu, J., Shi, L., Lin, W. P., & Tsai, B. Y. (2020). Customer segmentation based on RFM model and K-Means clustering. Sustainability, 12(5), 1970. https://doi.org/10.3390/su12051970

Wang, X., Wang, Z., Sheng, M., Li, Q., & Sheng, W. (2021). An adaptive and opposite K-means operation based memetic algorithm for data clustering. Neurocomputing, 437, 131–142. https://doi.org/10.1016/j.neucom.2021.01.056

Wu, Q., Benlic, U., & Hao, J.-K. (2021). Responsive threshold search based memetic algorithm for balanced minimum sum-of-squares clustering. Information Sciences, 569, 184–204. https://doi.org/10.1016/j.ins.2021.04.014

Xu, R., & Wunsch, D. (2009). Clustering. Wiley-IEEE Press.