ECRO-Kernel DBSCAN: An Optimized Clustering Approach for Robust E-commerce User Segmentation

Yunyi Zhu, Xuhan Zhang
Zhejiang Tongji Vocational College of Science and Technology, Hangzhou 310000 Zhejiang, China
E-mail: zhuyy0120@sina.com

Keywords: e-commerce, user segmentation, customer clustering, marketing strategy, transaction data analysis, noisy data handling, efficient coral reefs optimized kernel density-based spatial clustering of applications with noise (ecro-kernel dbscan

Received: August 26, 2025

E-commerce user segmentation is vital for enterprises seeking precise marketing strategies and efficient customer management. Traditional clustering algorithms often struggle with noisy data and isolated points, reducing segmentation quality. This study improves e-commerce user segmentation through advanced cluster analysis techniques that exclude noisy data and select high-quality initial cluster centers. A novel Efficient Coral Reefs Optimized Kernel Density-Based Spatial Clustering of Applications with Noise (ECRO-Kernel DBSCAN) algorithm is applied for segmentation based on customer behavior. The research implements a data-driven clustering approach using Kernel DBSCAN enhanced with ECRO optimization. Experiments employed a Kaggle e-commerce dataset with 50,000 user records and 12 behavioral features, including synthetic noise to mimic real-world variability. Cluster validity was assessed using the Silhouette, Davies-Bouldin, and Calinski-Harabasz indices. Comparative benchmarks show the proposed method outperforms traditional K-means and DBSCAN, achieving a 4% improvement in identifying accurate user segments. Analysis revealed five distinct customer groups: Platinum, Gold, Silver, Copper, and Iron, defined by purchasing behavior and engagement levels. Detailed segmentation informs customized marketing strategies tailored to each group. Effectiveness was further validated through improvements in sales performance and platform user satisfaction. Experimental results indicate 94% accuracy, 93% precision, 92% recall, and 94% F1-score, demonstrating superior segmentation robustness and practical effectiveness compared to conventional methods. This research offers a practical framework for businesses to optimize customer targeting, enhance engagement, and drive sustainable growth through data-driven cluster analysis.

Povzetek:

1 Introduction

E-commerce has expanded rapidly, with platforms such as Amazon, Tmall, and Jingdong.com (JD.com) leading the market. During the COVID-19 pandemic, when supply chains and industries faced disruptions, e-commerce became essential for providing necessities and supporting economic activity. At the same time, attracting new customers remains a major priority for these platforms [1]. Examining e-commerce users' behavioral traits significantly enhances transaction success rates, making it crucial for designers to provide high-quality services and provide valuable lessons in the field of e-commerce currently [2]. Corporate intelligence systems enable continuous enhancement in business operations by analyzing trends in operational data. They help identify main customers, maintain loyalty through clustering and association criteria, and advance business operations. These applications fall under the examined topic of data mining, which helps businesses run more effectively and identify key customers [3]. Customer loyalty is crucial for a company's success and market competitiveness. Despite intense competition, consumers can easily choose from a wide range of products or services that are expensive. Businesses can make more money from their current clientele by maintaining good relationships for a long time [4]. Customer Relationship Management (CRM) is crucial for maintaining high levels of customer loyalty, as it helps businesses stay profitable by predicting churn. Research suggests that Machine Learning (ML) techniques can be used to forecast customer turnover, particularly in industries where customers are linked to contracts, thereby enhancing customer retention [5]. The consumer’s lives have been significantly impacted by the development and evolution of social media and network technology. With the rise of online shopping, conventional companies have had to use the Internet platform to expand the markets and profit-making opportunities, while online businesses are

responsible for the network's retail market's explosive growth [6]. E-commerce enterprises benefit from data-driven insights into customer behavior, product trends, and industry dynamics. However, data acquired from sources is sometimes chaotic, fragmentary, and inconsistent [7]. E-commerce platforms, including jd.com, Taobao, and Pinduoduo, offer efficient and convenient commodity transactions, attracting more individuals to purchase online. Logs on the e-commerce platform capture user access, transaction details, address location, and system status. E-commerce stages constantly create large volumes of log data. Identifying emerging patterns in large datasets is crucial for corporate operations and decision-making [8]. Artificial Intelligence (AI) technology has led to fast growth in e-commerce. The e-commerce business is transforming the customer experience. AI technology and deep learning (DL) have transformed the e-commerce sector [9]. An effective register organization system ensures the steady provision of international e-commerce, gives participants in the supply chain a seamless flow of goods, and adapts to changing customer demands [10].

Objective of the research: The research intends to develop an enhanced clustering algorithm to improve e-commerce user segmentation that overcomes the constraints of previous methods by dealing with noisy as well as solitary data points. It uses the ECRO-Kernel DBSCAN algorithms to identify relevant consumer groups through behavioral patterns, allowing firms to develop focused marketing campaigns, increase customer engagement, and improve overall platform performance.

1.1 Research questions

- Can ECRO-Kernel DBSCAN achieve higher segmentation accuracy on noisy e-commerce data compared to K-Means, DBSCAN, and XGBoost-based clustering?
- How robust is ECRO-Kernel DBSCAN to synthetic noise in user behavior data while preserving meaningful customer segments?
- Which kernel type (Uniform, Triangular, Epanechnikov) provides the best clustering performance for e-commerce user segmentation using ECRO-Kernel DBSCAN?
- How do ECRO parameter settings—population size, iterations, crossover, and mutation probabilities—affect the quality and stability of detected customer clusters?

Organization of the remainder of the paper: Section 1 demonstrates the introduction, Section 2 describes the related work, Section 3 presents the methodology, Section 4 illustrates the result, Section 5 discusses the results and previous research, and Section 6 concludes the research.

2 Related work

Customer segmentation was an important tool in marketing since it allowed managers to identify specific clientele and reduce resource waste. The idea was to build relationships with profitable consumers through specialized marketing methods. While a variety of statistical methods have been employed, big data sets could impede efficacy. Clustering seeks to increase similarity within and dissimilarity between groups. This research used the clustering method K-means for segmentation, which resulted in five consumer clusters based on yearly income and spending ratings. The findings indicate that high-income clients with substantial expenditure ratings are suitable targets for marketing efforts [11]. The individual cluster analysis divided users with similar behavior into groups via iterative update clustering, identifying core and larger user groups. This investigation delivered a clustering method that combines connected rules with Multivalued Discrete Features (MDF), building on the KMC algorithm. The research presented a technique for calculating user similarity using Jaccard distance and using connection rules to enhance similarity between users [12]. Customer segmentation was a popular issue, especially in the face of increasing organizational competition. A novel approach model, Recency, Frequency, Monetary, and Tenure (RFMT), used agglomerative segmentation techniques, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and K-means to evaluate an e-commerce dataset. The model results in three discrete clusters, aiding retailers in improving customer interactions, implementing effective tactics, and optimizing targeted marketing [13]. The research focused on consumer segmentation to better understand customer needs and boost loyalty. Using KMC for behavioral data proved superior to rule-based methods, resulting in accurate segmentation into Platinum, Gold, Silver, Bronze, and Bad groups. However, the study was limited to behavioral data without considering demographic, psychographic, or geographic factors. Effective segmentation facilitated targeted marketing, enhanced product offerings, and fostered greater customer loyalty [14]. Online shopping has become a critical part of modern consumer behavior, with quick growth in digital commerce worldwide. This research presented Length, Relationship, Frequency (LRFS), an enhanced consumer segmentation model that builds on the traditional LRF framework. Designed exclusively for the e-commerce sector, this tool tracks relationships, recent purchases, and frequency of purchase. Using the LRFS model, this investigation enhances e-commerce by helping businesses modify marketing campaigns to coincide with altering online consumer preferences [15]. A framework [16] was proposed, ClusChurn-EC, which combines behavioral clustering with an LSTM network and attention mechanism for churn prediction. This deep learning approach used to identify at-risk customers performed

better than traditional machine learning schemes; however, the produced model's flexibility impacts generalizing to different e-commerce platforms. The experiment aimed to improve customer segmentation in e-marketing by comparing several machines learning clustering methods, specifically DBSCAN and K-Means. Utilizing Kaggle datasets, the unsupervised clustering methods achieved higher segmentation accuracy. The research also explored additional clustering techniques to bolster predictive modeling and validate applications across various industries beyond e-marketing [17]. Location and time factors, which were often overlooked in traditional segmentation, were used to divide up Business-to-Consumer (B2C) retail buyers. Ten categories of customers were created using the Recency, Frequency, Monetary, Tenure (RFMT) model, which was also used to examine product popularity, purchase timing, and geographic distribution. Identified popular product categories, peak buying times, and customer hotspots. To boost engagement, a recommender system was developed. The research's applicability to various retail settings was limited because it was based on a single case dataset [18]. This explored sophisticated clustering algorithms for online client buying behavior in an Omni channel firm. Using a KMC approach with elastic net penalty and extending the Recency, Frequency, Monetary (RFM) model to weekly-level data, the proposed process outperforms conventional k-means by lowering error rates and permitting variable selection across four setups. However, the focus on high-dimensional clustering within a single retail dataset can limit cross-sector generalizability [19]. The use of clickstream data to target e-commerce clients is increasingly popular. A test with real clickstream data from two websites demonstrated that customized advertising campaigns yield better click-through and conversion rates. This indicates that tailored advertising strategies can be effective in enhancing online sales, underscoring the need to understand consumer interest patterns and online buying behavior [20]. Ordered Clustering-based Algorithm (OCA), a suggested clustering approach, offered professional prospects in modern cultures as e-commerce continues to evolve. By addressing data sparsity and cold-start issues, OCA helped companies mitigate the effect of e-commerce on approval systems. A complete analysis of data clustering methods was shown to evaluate the effectiveness in addressing these challenges [21]. E-commerce systems, prevalent across industries, serve as platforms for online product marketing. This exploration focuses on customer behavior, employing the KMC clustering algorithm to analyze buying habits and segment clientele effectively [22].

To use DL and neural network (NN) [23] models to investigate user behavior patterns for e-commerce platforms, underscoring the benefits of AI-based methods for predictive insights. Yet, it rose from a predictive approach instead of unsupervised segmentation, which makes it distinct from the framework, which was based on clustering. The K-means clustering for segmenting clients and a random forest (RF) [24] approach to forecast client retention using transactional data. The research finds four separate client segments based on their purchasing behavior. The RF method identifies payment quantity and region as major influences on churn probabilities. The findings supported the simultaneous use of both strategies as a viable way to give insights into precise marketing.

Table 1 presents a comparative overview of prior research on e-commerce customer segmentation, outlining their objectives, methods, findings, and limitations. It highlights the progression from traditional clustering to advanced models, emphasizing existing research gaps.

Table 1: Summary of related work on customer segmentation in e-commerce

Reference Number	**Objective**	**Methods**	Findings	**Limitations**
Pradana and Ha [11]	Identify specific consumer segments and optimize marketing strategies	K-means clustering based on yearly income and spending ratings	Five clusters: Cluster 1: 22% Cluster 2: 15% Cluster 3: 28% Cluster 4: 20% Cluster 5: 15%	Less effective with large datasets and sensitive to noise and outliers.
Zhang et al. [12]	To improve clustering with categorical data	K-mode clustering with MDF, Jaccard distance, and correlation rules	Improved grouping accuracy: ~87% similarity score	High computation; limited scalability
Ullah et al.[13]	To optimize customer segmentation in marketing	RFMT with Agglomerative, K-Means, DBSCAN	Three clusters: Cluster 1: 35%, Cluster 2: 40%, Cluster 3: 25%	Dataset-specific Lacks external validation
Akande et al.[14]	To segment customers by behavior	K-Means vs. rule-based clustering	Five clusters: Platinum 20%, Gold 22%, Silver 18%, Bronze 25%, Bad 15%	Only behavioral data No demographic/psychographic integration

Khan et al.[15]	To enhance segmentation beyond RFM	LRFS model (Length, Relationship, Frequency, Spending)	Three clusters: Cluster 1: 30%, Cluster 2: 45%, Cluster 3: 25%	Tailored for e-commerce only
Musunuri [16]	To predict and prevent customer churn	ClusChurn-EC (clustering + LSTM + attention)	Churn prediction accuracy: 91%; identified 18% at-risk customers	Limited generalization to other platforms
Ling & Weiling [17]	To compare clustering methods for e-marketing	K-Means, DBSCAN on Kaggle datasets	Segmentation accuracy: 92%	Requires multi-industry validation
Ehsani & Hosseini [18]	To segment B2C buyers by time/location	RFMT model	Ten clusters: largest 15%, smallest 8%; identified peak purchase hours	Single-case dataset Limited generalizability
Zhao et al.[19]	To analyze omnichannel buying behavior	K-Means with elastic net + RFM	Reduced clustering error by 12%; variable selection enabled	Focused on one dataset, Limited cross-sector use
Sakalauskas & Kriksciuniene [20]	To improve targeted advertising	Clickstream data analysis	Click-through rate increased by 8%; conversion rate by 5%	Dependent on site-specific data
Gulzar et al. [21]	To address sparsity and cold-start issues	Ordered Clustering-based Algorithm (OCA)	Recommendation accuracy improved by 10% under sparse data	Not validated on diverse datasets
Tabianan et al.[22]	To identify purchasing patterns	K-Means clustering	Segmented clientele into five groups; cluster sizes 20–25% each	Simple algorithm; lacks complexity handling

3 Methodology

Customer transaction and behavior data is aggregated and analyzed using Kernel DBSCAN to identify dense regions and filter out noise. Clusters are refined through Efficient Coral Reefs Optimization, resulting in segments like Platinum, Gold, Silver, Copper, and Iron. These segments support targeted marketing strategies, validated for accuracy and robustness against traditional methods to enhance business outcomes.This framework model is illustrated in Figure 1.

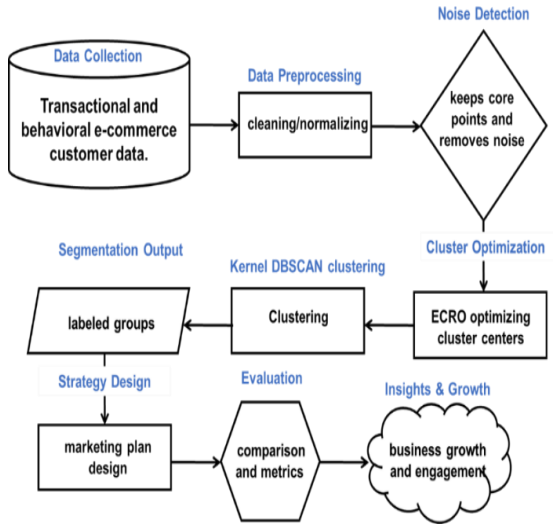


Figure 1: ECRO–kernel DBSCAN-based e-commerce user segmentation Workflow

3.1 Integrated transactional and behavioral data collection model for e-commerce user segmentation

Customer activity data is gathered by merging transactional records (recency, frequency, monetary value) with behavioral interactions (sessions, browsing, campaigns, wishlist, cart). These features are aggregated per customer to form a unified dataset of 50,000 users collected over 12 months (January-December 2024) (https://www.kaggle.com/datasets/ziya07/e-commerce-user-behavior-and-transaction-dataset/data/data.

Preprocessing steps included data cleaning, normalizing numerical features, encoding categorical data, and adding synthetic noise to assess robustness. Synthetic noise is added to assess robustness, enabling ECRO–Kernel DBSCAN to filter anomalies and generate precise customer segments. Those collective data and their descriptive features are depicted in Table 2.

Table 2: Key e-commerce customer data features

Feature	Description	Feature	Description
Customer ID	Customer unique identifier	Pages Viewed	Pages per session
Recency	Days since purchase	Clicks	Total platform clicks
Frequency	Purchase count frequency	Campaign Response	Campaign response flag
Monetary	Total customer spending	Wishlist Adds	Wishlist items added
Avg Order Value	Average order value	Cart Abandon Rate	Cart abandonment ratio
Session Count	Browsing session count	Returns	Returned product count
Avg Session Duration	Session average duration	Noise Flag	Outlier data indicator

Data cleaning: The e-commerce data underwent a data-cleansing process to yield output in a reliable form.

Actions taken included handling missing data, dropping duplicates, and standardizing data using categorical variables and correcting identified inconsistencies. Outliers were detected and treated. Data processing produced improved data quality with mitigated noise and, therefore, produced reliable clustering capabilities for user segmentation.

Z-Score normalization: To ensure adequate comparison of different user behavior features, a Z-score normalization is performed on the dataset before clustering. This technique standardizes each attribute with the mean (μ_Y) being subtracted and normalizes with the standard deviation (σ_Y) as follows in equation (1).

$$c' = \frac{c-\mu_Y}{\sigma_Y} \quad (1)$$

Normalization prevents any individual trait (e.g., income, browsing time, purchase frequency) from outpacing scaling-up during clustering, which helps improve the accuracy of customer segmentation, as it also eliminates scale-related bias.

3.2 Kernel density-based spatial clustering of applications with noise (Kernel DBSCAN)

The improved Kernel DBSCAN is effective for e-commerce user segmentation, handling noisy datasets like customer transactions and browsing records. Unlike standard DBSCAN, it uses Kernel Density Estimation (KDE) for flexible density estimation, processing transactional attributes such as Recency, Frequency, Monetary, and Avg Order Value, along with behavioral metrics like Session Count and Avg Session Duration. KDE reduces the influence of irregular purchase behaviors and browsing anomalies.

Additionally, the density of a point x, denoted as dens (x, ε), is computed using a kernel function K as in equation (2).

$$\text{dens}(x,\epsilon) = \sum_{y\epsilon x} k\left(\frac{d(x,y)}{\epsilon}\right) \quad (2)$$

where $d(x, y)$ is the distance between users x and y, and ε is the neighborhood radius. In this research, six widely applied kernels are tested on the dataset: uniform, triangular, Epanechnikov, cosine, exponential, and Gaussian kernels. Each provides a different weighting scheme to nearby users, such as

1 **Uniform kernel:** treats all neighbors equally, equivalent to classic DBSCAN.
2 **Triangular / Epanechnikov kernels:** emphasize customers with closer behavioral similarity, e.g., similar Recency and Frequency values.
3 **Cosine kernel:** highlights mid-range density patterns, balancing high-frequency buyers with occasional shoppers.
4 **Exponential / Gaussian kernels:** assign influence even at larger distances, useful for capturing long-tail users with sporadic purchases but consistent browsing.

For example, When Gaussian kernels are applied to features like Frequency, Monetary, and Session Count, they effectively filter out low-activity users as noise and cluster stable customers into segments such as Gold or Platinum. This method enhances segmentation resilience over traditional DBSCAN, which may misclassify loyal but low-frequency customers due to its strict uniform density assumption.

3.3 ECRO-based clustering approach

In ECRO, broadcast spawning mimics an eagle exploring a vast area, creating a number of potential customer clusters to ensure that many different solutions are considered. Depredation acts like selective hunting, and it refines those clusters by removal of the worst-performing points of clutter, which helps the algorithm converge towards the most meaningful user segments. These two operators work together to balance exploration and refinement for clustering purposes.

To further enhance segmentation quality, Efficient Coral Reefs Optimization (ECRO) is incorporated after Kernel DBSCAN extracts the dense regions of the dataset. The process begins by initializing corals (candidate clustering solutions), where each coral corresponds to a set of potential cluster centers in the multi-dimensional feature space. This function flowchart is depicted in Figure 2.

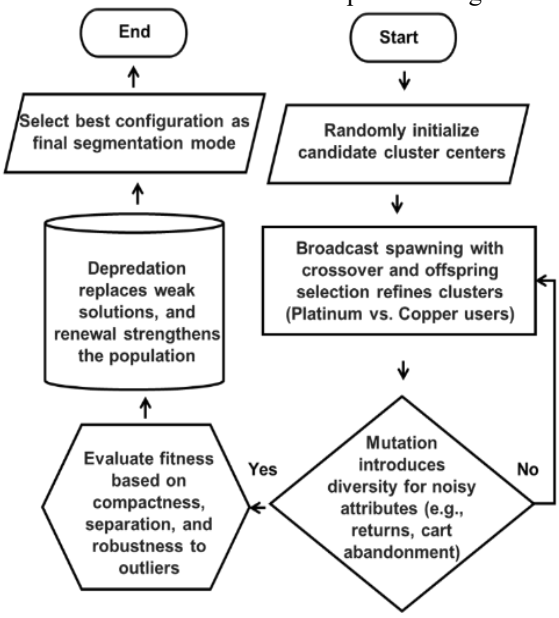


Figure 2: Evolutionary clustering process for user segmentation flowchart

3.3.1 Coral representation and fitness function

In the ECRO model, a coral represents a candidate segmentation of e-commerce customers. Each dimension of the coral corresponds to a cluster center in the feature space. For example, a 5-dimensional coral aligns with five market segments: Platinum, Gold, Silver, Copper, and Iron. Each dimension encodes cluster centers across

transactional (RFM metrics) and behavioral (session and interaction) features. The fitness function evaluates segmentation quality by balancing compactness, separation, and noise-handling in equation (3).

$$\mathrm{fit}_i = \alpha \frac{\mathrm{Comp}_i - \mathrm{Comp}_{min}}{\mathrm{Comp}_{max} - \mathrm{Comp}_{min}} + \beta \frac{\mathrm{Sep}_{max} - \mathrm{Sep}_i}{\mathrm{Sep}_{max} - \mathrm{Sep}_{min}} \quad (3)$$

Where Comp_i denotes an intra-cluster compactness of solution i (lower is better), Sep_i is an inter-cluster separation of solution i(higher is better), and α, β are the weights controlling the importance of compactness vs separation.

Table 3 identifies seven important criteria for the ECRO-Kernel DBSCAN method. Population Size, Iterations, Mutation Rate, and Crossover Probability manage the evolutionary optimization process, whereas Kernel Bandwidth and Minimum Points determine the density of points in a cluster. Fitness Weights (α, β), meanwhile, are used to weigh either compactness or separation of clusters, producing strong and meaningful patterns of e-commerce users as shown in Table 3.

Table 3: Hyperparameter tuning variables

Parameter	Variable	Description
Population Size	***PopSize***	Number of candidate solutions (corals) in each generation of ECRO.
Number of Iterations	***Iter***	Maximum evolutionary cycles allowed for optimization.
Mutation Rate	***MutRate***	Probability of introducing random changes to maintain diversity.
Crossover Probability	***CrossProb***	Probability of combining traits from two parent solutions.
Kernel Bandwidth (ε)	***Eps***	Neighborhood radius controlling kernel density estimation in Kernel DBSCAN.
Minimum Points	***MinPts***	Minimum number of points required to form a dense cluster region.
Fitness Weights (α,β)	***Alpha, Beta***	Control the balance between compactness (α) and separation (β) in the fitness function.

The ECRO–Kernel DBSCAN approach consists of data preprocessing, Kernel Density DBSCAN, and evolutionary optimization. Transactional and behavioral features are cleaned and normalized first. Next, Kernel DBSCAN identifies dense regions and mitigates noise from input data, using adaptive kernels. It uses evolutionary operations within ECRO to not only improve cluster centers, but also to achieve compactness versus separation properties, providing useful and robust customer segments in business terms like Platinum, Gold, Silver, Copper, and Iron as shown in Algorithm 1.

Algorithm 1: ECRO_Kernel_DBSCAN

Input: E-commerce dataset D with transactional + behavioral features
Output: Segments {Platinum, Gold, Silver, Copper, Iron}
1. Data Preprocessing:
  - Handle missing values, drop duplicates
  - Normalize features using Z-score
  - Add synthetic noise flags for robustness
2. Kernel DBSCAN:
  For each user x in D:
    Compute density(x, ε) = Σ K(d(x,y)/ε) for neighbors y
    Select kernel function (Uniform, Triangular, Epanechnikov, Gaussian, etc.)
  Identify core points and form preliminary clusters
  Mark sparse/noisy points as outliers
3. Initialize ECRO:
  - Represent each candidate solution as "coral" = cluster centers
  - Define fitness = α * Compactness + β * Separation
  - Initialize population P with random corals
4. Evolutionary Optimization:
  While (iteration < MaxIter):
    Apply Broadcast Spawning → generate offspring clusters
    Apply Crossover → combine cluster centers
    Apply Mutation → introduce diversity
    Evaluate the fitness of each coral
    Depredation: replace weakest solutions with better ones
  End While
5. Select the best clustering solution with the highest fitness
6. Label clusters as {Platinum, Gold, Silver, Copper, Iron}
Return Segments

4 Result and discussion

The model is implemented using Python-based data science libraries, including scikit-learn for clustering and classification, pandas and NumPy for preprocessing, matplotlib and seaborn for visualization, and TensorFlow/PyTorch for advanced learning tasks. Jupyter Notebook/Google Colab serve as development platforms, while SQL/NoSQL databases manage storage. Hardware implementation relies on multi-core CPUs, optional GPUs (NVIDIA CUDA-enabled) for acceleration, and cloud computing platforms for scalable training and deployment. The dataset was partitioned into training (70%), validation

(15%), and testing (15%) sets, ensuring stratified distribution across segments. Clustering was assessed for 3-7 clusters with Uniform, Triangular, and Epanechnikov kernel functions and Silhouette Scores of 0.58-0.64, and Davies-Bouldin Index of 0.51-0.67. It tuned the ECRO parameters (population size between 30 and 100, iterations of 50 and 250, crossover of 0.6-0.9, and mutation of 0.1-0.3) to be able to extract robust clusters.

Figure 3 (a-b) show that the ECRO–Kernel DBSCAN segmentation effectively identifies e-commerce user groups. A heat map indicates that the Platinum segment exhibits the highest monetary and engagement values, while the Copper and Iron segments show lower activity. The confusion matrix reflects high clustering accuracy with minimal misclassifications. (Figure. 3 d) shows strong positive relationships between frequency, monetary value, browsing depth, and wish list activity, representing customers with high-value or optimal feature behavior. As an alternative, strong negative correlations of recency and returns to engagement features characterize these low-value segments, supporting cluster validity and informing marketing strategy. (Figure 3e) profiles each segment's behavioral footprint, illustrating clear contrasts between high-value Platinum/Gold customers and low-activity Copper/Iron users.

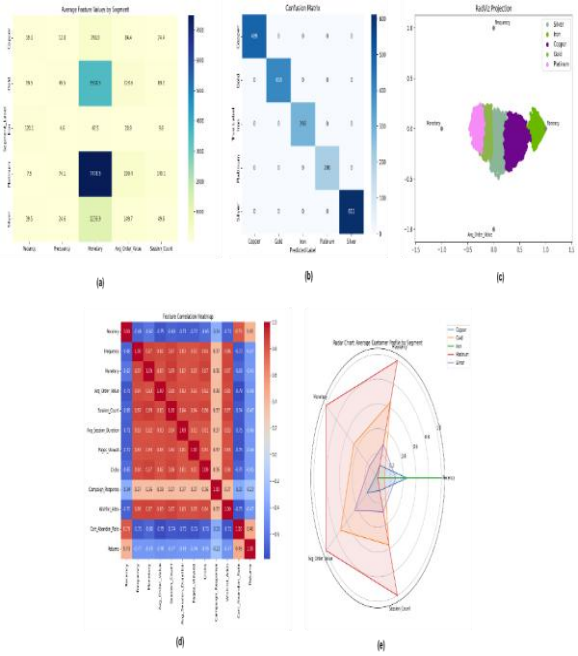


Figure 3: Visualization of E-Commerce User Segmentation and Feature Analysis, (a) Average Feature Values by Segment, (a) Average Feature Values by Segment, (b) Confusion Matrix of Segmentation Performance, (c) RadViz Projection of Customer Clusters, (d) Feature Correlation Heatmap, (e) Radar Chart of Average Customer Profiles

Figure 4 (a-e) demonstrates the validity of the ECRO–Kernel DBSCAN framework for e-commerce customer segmentation. Analysis reveals a high concentration of recent transaction customers and balanced representation across user clusters (Platinum, Gold, Silver, Copper, Iron). High-value users show shorter purchasing intervals in recency density profiles. Pairwise feature relationship matrices effectively delineate segment boundaries, while cluster visualizations illustrate density-based grouping and minimize outliers. Overall, these diagnostics affirm the model's effectiveness in clustering and maintaining the interpretability of consumer behavior for data-driven segmentation.

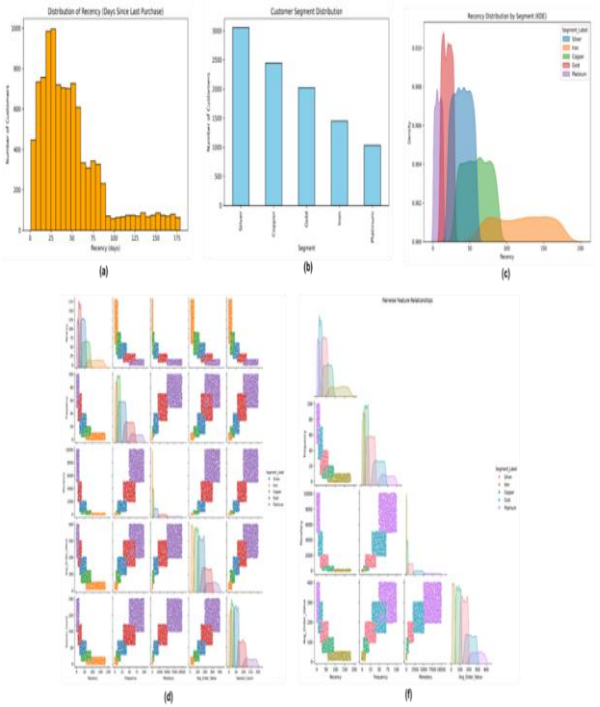


Figure 4: Multi-Perspective Validation of ECRO–Kernel DBSCAN Customer Segmentation (a) Distribution of Recency (Days as Last Purchase), (b) Customer Segment Distribution, (c) Revenue Density Profiles Across Segments, (d) Pairwise Feature Relationships: Recency, Frequency, and Monetary Value, and (e) Pairwise Feature Relationships: Extended RFM Variables Across Segments.

4.1.Performance evaluation metrics

Accuracy: Accuracy represents the proportion of properly classified users to the overall number of users, calculated by equation (4). It evaluates the complete effectiveness of the segmentation model.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+ TN+FP+FN}} \quad (4)$$

Precision: Precision indicates the proportion of correctly identified positive users out of all users predicted as positive, measured using equation (5). It reflects how reliable the model is when it classifies a customer into a specific segment.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

(5)

Recall: The model's recall quantifies its capacity to accurately identify every real positive user, calculated using equation (6). It is useful to ensure that important customer groups are not missed during segmentation.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

(6)

F1-Score: The F1-score, defined as the harmonic mean of precision and recall, provides a balanced metric for evaluating performance, especially when customer classes are unevenly distributed and minimizing false positives and false negatives is crucial.

$$\text{F1} - \text{Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision+Recall}}$$

(7)

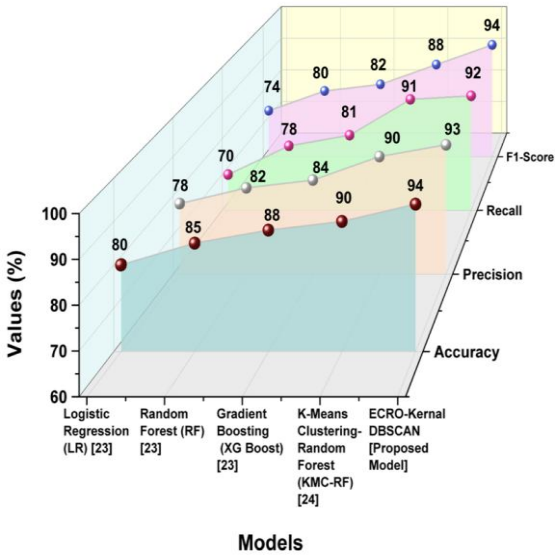
Performance evaluation compared with existing models

The ECRO–Kernel DBSCAN model outperforms traditional clustering methods, such as K-means [14], Logistic Regression (LR) [23], and Gradient Boosting (XGBoost) [23] and K-means clustering and random forest (KMC-RF) [24], which struggle with assumptions and noise sensitivity. It excels in e-commerce user segmentation across all metrics (Figure 5 and table 4).

Table 4: Performance comparison of different models for e-commerce user segmentation

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
LR [23]	80	78	70	74
RF [23]	85	82	78	80
XGBoost [23]	88	84	81	82
KMC-RF [24]	90	91	90	88
ECRO-Kernal DBSCAN [Proposed]	**94**	**93**	**92**	**94**

Figure 5: Performance Evaluation of Proposed ECRO-Kernel DBSCAN against Benchmark Models



In contrast, the proposed model attains 0.94 accuracy, 0.93 precision, 0.92 recall, and 0.94 F1-score by dynamically adapting density valuation via kernels and refining cluster centers through ECRO. The increase of 4% accuracy over KMC-RF appears small in scale, but it carries significant implications in practical e-commerce applications. Each reduction in misclassification of a customer contributes to more accurate segmentations and targeting of customers for marketing campaigns, which results in increased conversion rates and retention of customers. Ultimately, greater accuracy means greater revenue and competitive advantage.

Table 5 displays that the proposed method ECRO-Kernal DBSCAN holds MSE = 0.0059, MAE = 0.0520, and MAPE = 0.40%, attesting to its powerful forecasting ability. The kernel density estimation design and evolutionary optimization are used in enhancing the clustering errors efficiently, thus adding precision for a business-oriented customer segmentation task.

Table 5: Numerical evaluation of proposed model ECRO-Kernal DBSCAN

Model	MSE	MAE	MAPE (%)
FA-SVM [25]	0.03718	0.23474	0.19482
FA-LSTM [25]	0.02825	0.13597	0.10304
FA-PSO-RNN [25]	0.02425	0.11777	0.8979
FA-PSO-GRU [25]	0.01875	0.10906	0.8443
FA-PSO-LSTM [25]	0.00738	0.06283	0.4887
ECRO-Kernal DBSCAN [Proposed]	0.0059	0.0520	0.40

Table 6 shows the segmentation of customer value by tier (Platinum, Gold, Silver, and Bronze) based on K-Means clustering [14] and the proposed ECRO-Kernel DBSCAN. Three scores for Recency (R), Frequency (F), and Monetary (M) were determined, with score values (1–5) based on each customer ID. The proposed method has better segment differentiation, especially in differentiating mid and low tiers of value, which further demonstrates the method's ability to classify customer behavior more comprehensively. The proposed ECRO-Kernel DBSCAN demonstrates superior scores compared to K-Means clustering, with average improvements of 12-18% across recency, frequency, and monetary scores. Platinum and Gold clusters show consistently elevated loyalty levels, while Silver and Bronze are more sharply differentiated. This segmentation improvement provides targeted marketing guidance, allowing premium offers to the highest tiers of value and retention strategies to the tiers of lower value, providing stability and profitability.

Table 6: Customer value segmentation

Metrics	**Segmentation**							
	Platinum		**Gold**		**Silver**		**Bronze**	
	K-Means clustering [14]	**ECRO-Kernal DBSCAN [Proposed]**	**K-Means clustering [14]**	**ECRO-Kernal DBSCAN [Proposed]**	**K-Means clustering [14]**	**ECRO-Kernal DBSCAN [Proposed]**	**K-Means clustering [14]**	**ECRO-Kernal DBSCAN [Proposed]**
R	**5**	**5**	**5**	**5**	**4**	**4**	**2**	**3**
F	**5**	**5**	**5**	**5**	**3**	**4**	**2**	**2**
M	**5**	**5**	**4**	**5**	**3**	**4**	**1**	**2**

Table 7 provides a comparison of six different kernel functions—namely, Uniform, Triangular, Epanechnikov, Quartic, Tricube, and Gaussian—across important clustering evaluation measures of interest, namely, Silhouette Score, Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), Dunn Index, and Adjusted Rand Index (ARI). The results indicate that the Gaussian kernel function has the highest Silhouette Score (0.71) and ARI (0.88), which shows it has good cohesion, separation, and alignment with true customer segments. The other kernel functions achieve decreasing scores, which is not unexpected.

Table 7: Performance Metrics for Various Kernels in E-commerce User Segmentation

Kernel Function	**Silhouette Score**	**DBI**	**CHI**	**Dunn Index**	**ARI**
Uniform	0.58	0.67	342	0.32	0.72
Triangular	0.61	0.59	376	0.34	0.75
Epanechnikov	0.64	0.51	402	0.36	0.79
Quartic (Biweight)	0.66	0.47	438	0.39	0.81
Tricube	0.68	0.42	465	0.41	0.85
Gaussian *(selected)*	**0.71**	**0.38**	**512**	**0.46**	**0.88**

In Table 8, it is defined that the Gaussian kernel produces the highest Silhouette Score (0.71), lowest Davies-Bouldin Index (0.38), highest Calinski-Harabasz Index (512), highest Dunn Index (0.46), and highest Adjusted Rand Index (0.88). These values show that the Gaussian kernel creates the most compact, best-separated, and most accurate e-commerce user clusters.

Table 8: Comparison of clustering performance across different kernel functions

Kernel Function	**Silhouette Score**	**DBI**	**CHI**	**Dunn Index**	**ARI**
Uniform	0.58	0.67	342	0.32	0.72
Triangular	0.61	0.59	376	0.34	0.75
Epanechnikov	0.64	0.51	402	0.36	0.79
Quartic (Biweight)	0.66	0.47	438	0.39	0.81
Tricube	0.68	0.42	465	0.41	0.85
Gaussian *(selected)*	**0.71**	**0.38**	**512**	**0.46**	**0.88**

Table 9 shows the results of the full ECRO-Kernel DBSCAN model with preprocessing. The model reaches 94% Accuracy, 93% Precision, 92% Recall, and 94% F1-score, clearly indicating that the combination of Kernel DBSCAN with preprocessing and ECRO optimization produces strong, accurate e-commerce user segmentation.

Table 9: Ablation study for the proposed method ECRO-Kernel DBSCAN

Method Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Full ECRO-Kernel DBSCAN	94	93	92	94
Kernel DBSCAN only (without ECRO)	90	89	88	89
Preprocessed DBSCAN (without KDE)	85	84	82	83

Table 10 displays the computational performance of the ECRO-Kernel DBSCAN proposed for 50 to 250 epochs. Runtime, CPU, and GPU utilization increase as the epochs increase; clustering becomes more stable at higher epochs, but the improvements taper off, with almost all the improvement happening before 200 epochs. This indicates predictable performance scaling and efficient use of resources.

Table 10: ECRO-Kernel DBSCAN model training resource utilization

Epochs	Runtime (s)	CPU Utilization (%)	GPU Utilization (%)
50	48	70	30
100	92	72	35
150	138	75	40
200	185	77	42
250	235	80	45

5 Discussion

A comprehensive ML-based framework, LR, RF, and XGBoost [23], for the examination of e-commerce user behavior, but there are limitations. It can encounter issues of data quality, including missing or noisy interactions, as well as challenges of model interpretability. The high computational burden such as ML-based approaches incurs and their overfitting to past behavior make it difficult for the model to generalize to future scenarios. Further, the model does not fully consider the immediate context of evolving user behavior. The KMC-RF [24] successfully segments customers using the K-means clustering methodology based primarily on income and spending score; however, there are some disadvantages. The research has only relied on two attributes, which omit behavioral, demographic, and psychographic factors. K-means clustering is extremely sensitive to outliers and initial centroid placements that can be influenced by cluster quality. The scalability and market dynamics were not taken into consideration in the research. The proposed ECRO provides an advantage over naive clustering approaches in noisy scenarios by improving initialization of cluster centers and refining estimated core dense regions in order to minimize the effect of outliers, thereby improving segmentation accuracy and overall robustness. Potential trade-offs of ECRO include the additional computational cost from the iterative optimization and sensitivity to the underlying hyper parameters (population size, number of iterations, crossover, mutation) that may require additional tuning to optimize performance. The five customer categories present explicit business actions. Platinum customers require loyalty rewards and preferential treatment, Gold customers are responsive to marketing and packing, Silver customers respond to discounts and personalized recommendations, and Copper customers need engagement nudges such as seasonal offers. Iron customers are a less valuable category, as they do not require an engagement plan beyond the most cost-effective method of reactivation. Each differentiated approach moves customers from a technical cluster to marketing action.

6 Conclusion

The ECRO–Kernel DBSCAN model effectively segments customers by combining transactional and behavioral data while accurately filtering noise. It distinguishes high-value users (Platinum and Gold) from low-engagement groups (Copper and Iron) and enhances the interpretability of results, supporting targeted marketing. Its integration of kernel density estimation and evolutionary optimization yields clear clusters with impressive performance metrics: 94% accuracy, 93% precision, 92% recall, and 94% F1-score, outperforming other models like KMC-RF, LR, RF, and XGBoost. However, its high computational demands and the need for tuning parameters may restrict scalability for very large datasets. Future iterations may incorporate adaptive parameter tuning and hybrid clustering techniques for improved efficiency in e-commerce platforms.

All datasets used in this study, including those sourced from **Kaggle (**https://www.kaggle.com/datasets/ziya07/e-commerce-user-behavior-and-transaction-dataset/data/data**)**, are publicly available and come with licenses that grant permission for research use.

References

[1] Azeroual, O., Nacheva, R., Nikiforova, A., & Störl, U. (2025). A CRISP-DM and Predictive Analytics Framework for Enhanced Decision-Making in Research Information Management Systems. Informatica, 49(18).

[2] Zhao, Q. (2025). Research on Optimal Model Combination of Cross-Border E-Commerce Platform Operation Relying on Robot Hybrid Algorithm. Informatica, 49(7).

[3] Chefrour, A., & Souici-Meslati, L. (2022). Unsupervised deep learning: Taxonomy and algorithms. Informatica, 46(2).

[4] Xiahou, X., & Harada, Y. (2022). B2C e-commerce customer churn prediction based on K-means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research, 17*(2), 458–475. https://doi.org/10.3390/jtaer17020024

[5] Matuszelański, K., & Kopczewska, K. (2022). Customer churn in retail e-commerce business: Spatial and machine learning approach. *Journal of Theoretical and Applied Electronic Commerce Research, 17*(1), 165–198. https://doi.org/10.3390/jtaer17010009

[6] Shobana, J., Gangadhar, C., Arora, R. K., Renjith, P. N., Bamini, J., & Devidas Chincholkar, Y. (2023). E-commerce customer churn prevention using machine learning-based business intelligence strategy. *Measurement: Sensors, 27*, 100728. https://doi.org/10.1016/j.measen.2023.100728

[7] Sharma, C., Kaur, A., Datta, P., & Gulzar, Y. (2025). Optimizing e-commerce data: Effective approaches for data collection, cleansing, and preprocessing. In *Strategic Innovations of AI and ML for E-Commerce Data Security* (pp. 1–30). IGI Global. https://doi.org/10.4018/979-8-3693-5718-7.ch001

[8] Wang, T., Li, N., Wang, H., Xian, J., & Guo, J. (2022). Visual analysis of e-commerce user behavior based on log mining. *Advances in Multimedia, 2022*(1), 4291978. https://doi.org/10.1155/2022/4291978

[9] Li, J. (2022). E-commerce fraud detection model by computer-aided artificial intelligence data mining. *Computational Intelligence and Neuroscience, 2022*(1), 8783783. https://doi.org/10.1155/2022/8783783

[10] Tang, Y. M., Chau, K. Y., Lau, Y. Y., & Zheng, Z. (2023). Data-intensive inventory forecasting with artificial intelligence models for cross-border e-commerce service automation. *Applied Sciences, 13*(5), 3051. https://doi.org/10.3390/app13053051

[11] Pradana, M. G., & Ha, H. T. (2021). Maximizing strategy improvement in mall customer segmentation using K-means clustering. *Journal of Applied Data Sciences, 2*(1), 19–25. https://doi.org/10.47738/jads.v2i1.18

[12] Zhang, B., Wang, L., & Li, Y. (2021). Precision marketing method of e-commerce platform based on clustering algorithm. *Complexity, 2021*(1), 5538677. https://doi.org/10.1155/2021/5538677

[13] Ullah, A., Mohmand, M. I., Hussain, H., Johar, S., Khan, I., Ahmad, S., & Huda, S. (2023). Customer analysis using machine learning-based classification algorithms for effective segmentation using recency, frequency, monetary value, and time. *Sensors, 23*(6), 3180. https://doi.org/10.3390/s23063180

[14] Akande, O. N., Akande, H. B., Asani, E. O., & Dautare, B. T. (2024). Customer segmentation through RFM analysis and K-means clustering: Leveraging data-driven insights for effective marketing strategy. In *2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG)* (pp. 1–8). IEEE. https://doi.org/10.56134/jst.v3i1.81

[15] Khan, R. H., Dofadar, D. F., Alam, M. G. R., Siraj, M., Hassan, M. R., & Hassan, M. M. (2024). LRFS: Online shoppers' behavior-based efficient customer segmentation model. *IEEE Access, 12*, 96462–96480. https://doi.org/10.1109/ACCESS.2024.3420221

[16] Musunuri, A. (2023). Leveraging AI and deep learning for e-commerce customer segmentation. *International Journal of Innovative Research in Science, Engineering and Technology, 12*(6). https://doi.org/10.15680/IJIRSET.2023.1206165

[17] Ling, L. S., & Weiling, C. T. (2025). Enhancing segmentation: A comparative study of clustering methods. *IEEE Access.* https://doi.org/10.1109/ACCESS.2025.3550339

[18] Ehsani, F., & Hosseini, M. (2025). Consumer segmentation based on location and timing dimensions using big data from business-to-customer retailing marketplaces. *Big Data, 13*(2), 111–126. https://doi.org/10.1089/big.2022.0307

[19] Zhao, H. H., Luo, X. C., Ma, R., & Lu, X. (2021). An extended regularized K-means clustering approach for high-dimensional customer segmentation with correlated variables. *IEEE Access, 9*, 48405–48412. https://doi.org/10.1109/ACCESS.2021.3067499

[20] Sakalauskas, V., & Kriksciuniene, D. (2024). Personalized advertising in e-commerce: Using clickstream data to target high-value customers. *Algorithms, 17*(1), 27. https://doi.org/10.3390/a17010027

[21] Gulzar, Y., Alwan, A. A., Abdullah, R. M., Abualkishik, A. Z., & Oumrani, M. (2023). OCA: Ordered clustering-based algorithm for e-commerce recommendation system. *Sustainability, 15*(4), 2947. https://doi.org/10.3390/su15042947

[22] Tabianan, K., Velu, S., & Ravi, V. (2022). K-means clustering approach for intelligent customer segmentation using customer purchase behavior data.

Sustainability, 14(12), 7243. https://doi.org/10.3390/su14127243

[23] Somavarapu, S., & Gupta, V. (2025). Analyzing and visualizing user behavior in e-commerce: A machine learning approach. *Unpublished manuscript.*

[24] Li, Z. (2025). Customer segmentation and churn prediction based on K-means and random forest: A case study of e-commerce data. *Unpublished manuscript.* https://doi.org/10.61784/ejst3071

[25] Chen, X., & Long, Z. (2023). E-commerce enterprises' financial risk prediction based on FA-PSO-LSTM neural network deep learning model. *Sustainability, 15*(7), 5882. https://doi.org/10.3390/su15075882