iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** VIII **Month of publication:** August 2025

DOI: https://doi.org/10.22214/ijraset.2025.73583

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue VIII Aug 2025- Available at www.ijraset.com

Enhanced Customer Segmentation Using LRFSM Behavior Model and Multi-Algorithm Clustering Approaches

Sravan Kumar Mandeti
M.Tech, CSE department, UCEK, JNTU Kakinada, Andhra Pradesh, India

Abstract: ***Customer segmentation plays a critical role in optimizing personalized marketing and driving customer engagement in online retail. Building on the previously established LRFS (Length, Recency, Frequency, Spend) model, this study introduces an enhanced LRFSM framework by incorporating Monetary value as a fifth behavioral dimension. This expansion enables a deeper analysis of customer purchasing behavior and economic contribution. By applying advanced segmentation techniques and evaluating cluster quality through both internal validation scores and external classification metrics, this work ensures both statistical rigor and business relevance. Detailed profiling of each customer segment further provides meaningful insights into targeted strategy development. The results demonstrate that the inclusion of Monetary value not only improves segmentation precision but also supports the creation of scalable, adaptive models tailored to dynamic retail environments, bridging the gap between data science and actionable business outcomes.***

Keywords: ***Customer Segmentation, Behavioral Clustering, Unsupervised Learning methods, Retail Analytics, Clustering Evaluation.***

I. INTRODUCTION

In the rapidly evolving landscape of e-commerce, understanding customer behavior has become a critical factor in achieving long-term business success. With the growing volume of online transactions and customer interactions, businesses are now relying more than ever on data-driven techniques to segment their customers and deliver personalized experiences. Among these techniques, customer segmentation plays a vital role by categorizing customers into distinct groups based on shared characteristics, allowing for targeted marketing, improved customer service, and optimized resource allocation.

A commonly used method for customer segmentation is the RFM model, which evaluates customers based on Recency (how recently a customer made a purchase), Frequency (how often they make purchases), and Monetary value (how much they spend). Although this model provides valuable insights, it often lacks depth in capturing the full scope of customer behavior, particularly in long-term engagement and overall contribution. RFM's focus on short-term transactional metrics limits its effectiveness in understanding evolving customer relationships, especially in digital commerce environments where customer behavior is dynamic and multidimensional. To overcome these limitations, researchers have proposed the LRFS framework, which introduces two additional dimensions: Length (the duration of a customer's relationship with the platform) and Spend (the total cumulative amount spent over time). This enhanced model allows businesses to capture both the intensity and longevity of customer engagement, offering a more detailed behavioral profile than traditional methods. While LRFS has shown promising results when combined with clustering algorithms like KMeans, it still omits the individual Monetary value per transaction, which can provide further clarity into customer spending behavior. This research addresses that gap by introducing an extended model called LRFSM, which integrates five behavioral attributes: Length, Recency, Frequency, Spend, and Monetary. By combining both cumulative and average spending data with engagement metrics, the LRFSM model offers a more nuanced perspective on customer behavior. This expanded feature set enables more precise segmentation and allows businesses to differentiate between customers who spend frequently in small amounts and those who make fewer but higher-value purchases.

To evaluate the effectiveness of the LRFSM model, this study explores a variety of clustering algorithms beyond the commonly used KMeans. These include DBSCAN, which identifies clusters based on density; Spectral Clustering, which performs well with complex cluster shapes; KShape, suitable for time-series or sequence-based data; Agglomerative Clustering, which builds a hierarchy of clusters; Mini-Batch KMeans, a faster version of KMeans for larger datasets; and OPTICS, which can handle clusters of varying densities. Each algorithm is chosen to explore different structural assumptions and to assess which methods are most suitable for customer segmentation tasks using the LRFSM framework.

The evaluation of clustering results is conducted using both unsupervised metrics—such as Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index—and supervised metrics, including accuracy, precision, recall, and F1-score, with label alignment using the Hungarian algorithm for fair comparison. To provide business value, the study also includes cluster profiling, which interprets each cluster by analyzing the average behavior across the five features.

Additionally, the research considers future enhancements using deep learning-based clustering methods, such as Deep Embedded Clustering (DEC), which combine feature learning and clustering in a unified architecture. These models are expected to offer better performance with complex, high-dimensional behavioral data.

By expanding the LRFS model into LRFSM and evaluating it through multiple clustering approaches, this study aims to deliver a robust and adaptable framework for customer segmentation. The findings can support more effective marketing strategies, improve customer relationship management, and offer deeper insights into customer value across digital platforms.

II. RELATED WORK

Recent work by Ameen Al-Dubai et al. introduced the LRFS (Length, Recency, Frequency, Spend) model as an enhancement over traditional RFM segmentation, aiming to capture long-term customer engagement and cumulative spending behavior. Using KMeans clustering, the authors demonstrated improved segmentation accuracy and interpretability on e-commerce data. Their results showed that LRFS offers better customer insights compared to RFM and RF-only models. The study also emphasized the importance of cluster profiling and proposed future directions involving deep learning and hybrid clustering techniques [1].

One recent approach combined [2] spectral clustering with affinity propagation to create a dynamic and adaptive clustering algorithm suitable for complex behavioral datasets. This method automatically determines the optimal number of clusters and enhances the affinity matrix using eigenvector-based similarity transformations. The algorithm is particularly effective in identifying non-convex clusters and handling datasets with varying densities—common characteristics in customer behavior data. Its ability to model irregular structures without requiring a preset cluster number makes it well-suited for applications such as customer segmentation in digital environments. Another study tackled the problem of segmenting high-dimensional customer data,[3] where variable redundancy and correlation can hinder traditional algorithms. To overcome this, a regularized K-Means method was introduced that incorporates a penalty mechanism to reduce the influence of irrelevant or overlapping variables. This adjustment improves the clarity and separation of clusters, resulting in more robust and interpretable segmentation outcomes. The model is particularly beneficial when working with behavioral features like recency, frequency, and spend, which often exhibit correlation.

In the realm of long-term customer value analysis, one paper proposed a segmentation approach that integrates Customer Lifetime Value (CLV) components—such as retention rate, profit margin, and discount factors—into the clustering process. This model shifts the focus from short-term purchase behavior to long-term profitability, allowing businesses to identify customers who offer greater value over time. Such a perspective is highly relevant for strategic planning, enabling more targeted investment in retention and loyalty programs [4]. Another research effort enhanced [5] the classical RFM model by including multiple behavioral signals beyond transactional data. Interactions such as clicks, wishlist additions, and cart activity were incorporated using entropy-based weighting to capture more granular customer behavior. These behaviors were then clustered using an improved Self-Organizing Map (SOM) neural network, which performed well even with sparse data. The method showed strong potential in mobile commerce settings, where traditional RFM models may lack sufficient nuance. Outside the conventional retail domain, a study focused on clustering electricity consumers based on their responsiveness to tiered pricing in demand response programs. Unsupervised learning techniques were used to group consumers according to behavioral and usage patterns, helping energy providers design more effective incentive strategies. While the context differs, the underlying methodology offers valuable insights into behavioral segmentation and the design of adaptive, data-driven services—principles that are directly applicable to online retail and customer engagement [6]. Deep Embedded Clustering (DEC) was evaluated for its capability [7] to handle intensive care unit (ICU) data with mixed types (numerical and categorical). An adapted model, X-DEC, replaced the standard autoencoder with an X-shaped variational autoencoder to better manage mixed datatypes and optimize hyperparameters for cluster stability. The model was tested on two ICU datasets to assess internal and external validity. Results indicated that X-DEC generated more stable and generalizable clusters compared to DEC, highlighting its effectiveness for clinical decision support.

To address DEC's limitations with mixed data, [8] a modified deep embedded clustering framework was introduced, incorporating a soft-target update technique inspired by deep Q-learning to improve convergence stability. This approach effectively handled both numerical and categorical features without requiring transformation into a single format. Empirical evaluations on benchmark datasets showed that the proposed model consistently outperformed traditional clustering algorithms, achieving state-of-the-art results in standard clustering metrics.

A fast adaptive K-means subspace clustering (FAKM) model was developed to improve performance on high-dimensional datasets. By introducing an adaptive loss function and a feature selection mechanism that bypasses eigenvalue decomposition, FAKM efficiently performed clustering and feature extraction simultaneously. It was shown to be robust to noise and outliers and demonstrated superior computational efficiency and clustering accuracy on various benchmark datasets compared to conventional K-means-based models [9].

A hybrid approach was proposed to predict customer churn by combining statistical modeling and machine learning. The model used the Buy-Till-You-Die (BTYD) framework to estimate customer survival probabilities, followed by K-means clustering to segment customers into four behavioral types. [10] Machine learning algorithms were then applied to predict churn. Evaluation on two public e-commerce datasets showed strong performance recall, particularly in identifying high-value customers at risk of churning, proving the method's practical effectiveness for customer retention. An integrated method combining clustering and logistic regression was used to analyze online shopping behavior and forecast purchase decisions. [11] The dataset included both categorical and continuous features representing user interactions on e-commerce platforms. Cluster analysis grouped users based on usage characteristics like operating system and traffic source, while logistic regression identified significant factors impacting purchasing behavior within each group. The method revealed key differences in decision factors across clusters, supporting more personalized marketing strategies.

III. METHODOLOGY

This research adopts a multi-phased, data-driven methodology aimed at performing granular customer segmentation based on behavioral characteristics derived from transaction histories. The dataset utilized in this study comprises synthetically generated customer transaction records, modeled to resemble realistic e-commerce activity over a span of two years. It includes fields such as Customer_ID, Transaction_Date, and Monetary_Value, which collectively reflect customer purchasing behavior. Although synthetic, the dataset maintains structural and statistical fidelity to real-world customer datasets commonly encountered in online retail platforms. The preprocessing phase begins with the conversion of transaction timestamps into standardized datetime format. All records lacking a valid Customer_ID are excluded to ensure consistency in aggregation. Each unique Customer_ID is treated as a string for categorical grouping purposes. A reference date is then computed as one day after the most recent transaction in the dataset, which serves as a temporal anchor point for recency-based calculations.

Behavioral features are engineered through an extended framework referred to as LRFSM, which encompasses five key metrics: Length, Recency, Frequency, Spend, and Monetary value. These dimensions capture multiple aspects of customer engagement across time and spending behavior:

Length (L) represents the duration of the customer lifecycle, calculated as:

$$L = (\text{Last Transaction Date} - \text{First Transaction Date}) + 1$$

Recency (R) measures the time since the customer's most recent transaction:

$$R = \text{Reference Date} - \text{Last Transaction Date}$$

Frequency (F) reflects how often the customer has transacted:

$$F = \text{Count of Transactions for a given Customer\_ID}$$

Spend (S) refers to the total monetary amount spent by the customer:

$$S = \text{Sum of all Monetary Values associated with the Customer\_ID}$$

Monetary (M), the average spend per transaction, is computed as:

$$M = S / F$$

These features are further normalized using Z-score standardization, which ensures that each variable contributes equally during clustering, eliminating the bias caused by differing data scales.

To manage the complexity of high-dimensional data and enhance interpretability, Principal Component Analysis (PCA) is applied. This transformation projects the LRFSM feature space into two and three principal components that capture most of the variance. These reduced dimensions facilitate both visual cluster exploration and computational efficiency during modeling.

The core segmentation is performed using six unsupervised machine learning algorithms: KMeans, Agglomerative Clustering, Gaussian Mixture Models (GMM), DBSCAN, Spectral Clustering, and MiniBatch KMeans. Each algorithm contributes unique strengths: KMeans and MiniBatch KMeans offer scalable, centroid-based clustering for large datasets; Agglomerative Clustering provides a hierarchical perspective; GMM introduces probabilistic cluster assignment, allowing for uncertainty modeling; DBSCAN is adept at identifying clusters of arbitrary shapes and isolating noise; and Spectral Clustering leverages graph theory for separating non-convex clusters. This diverse selection ensures comprehensive exploration of the cluster space under various assumptions of data distribution.

Post-clustering, labels are appended to the original dataset, and clusters are visualized using 2D and 3D PCA scatter plots, enabling side-by-side comparisons across models. Each model's cluster arrangement is plotted using consistent color palettes and labeled legends to enhance interpretability.

To assess clustering performance, both intrinsic and supervised evaluation metrics are applied. Intrinsic metrics include:

- Silhouette Score:

$$S = (b - a) / \max(a, b)$$

where a is the average intra-cluster distance and b is the average distance to the nearest cluster.

- Davies-Bouldin Index (DBI): Lower values suggest better cluster separation.
- Calinski-Harabasz Index (CHI): Higher values indicate well-defined, compact clusters.

If external labels or aligned ground truths are available, supervised metrics such as Accuracy, Precision, Recall, and F1-Score are also computed to evaluate how well clustering aligns with actual customer classes.

Finally, cluster profiling is conducted by aggregating mean values of each LRFSM feature across clusters. This step enables behavioral interpretation of each customer segment, for instance, distinguishing between high-spending, frequent purchasers and newly acquired, infrequent users. These insights inform downstream applications such as personalized marketing, targeted promotions, and churn prediction strategies.

In summary, the proposed methodology presents a holistic and modular framework that integrates feature engineering, dimensionality reduction, unsupervised learning, and interpretability—offering a replicable and scalable approach to customer segmentation in dynamic online retail environments.

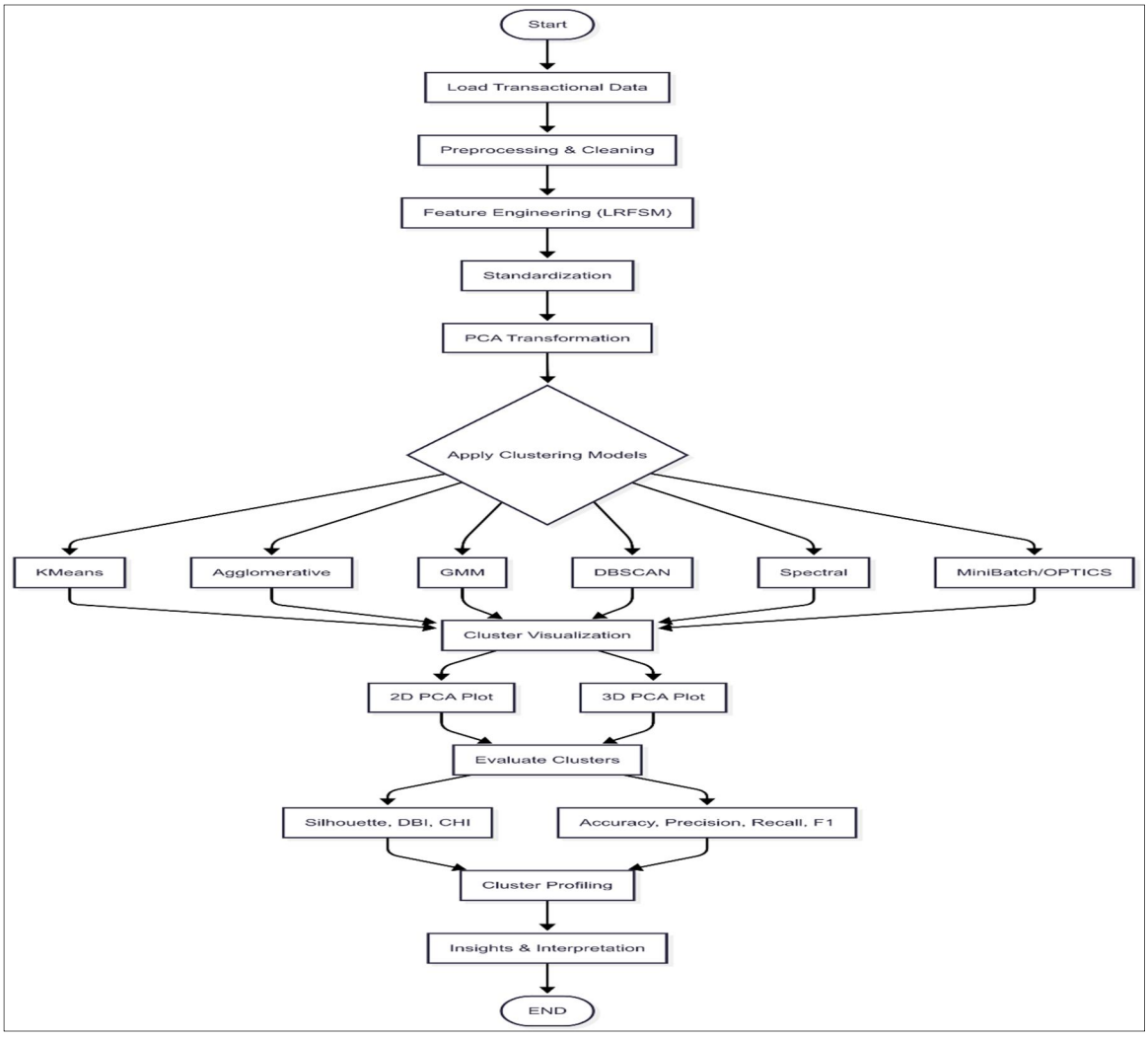


Figure 1: System Architecture for Behavioral Customer Segmentation

IV. RESULTS AND DISCUSSION

To evaluate the effectiveness of the proposed LRFSM-based customer segmentation framework, three clustering algorithms—DBSCAN, Spectral Clustering, and KShape—were applied to the standardized behavioral features derived from customer transaction data. The segmentation results were first visualized using 2D scatter plots (Figure 1), projecting Recency and Monetary dimensions. The DBSCAN output (Figure 1a) revealed a central dense cluster surrounded by several isolated noise points (labelled as -1), typical of its sensitivity to density thresholds. Spectral Clustering (Figure 1b) produced more uniformly distributed and well-separated clusters, reflecting its strength in graph-based space partitioning. In contrast, KShape Clustering (Figure 1c), which is tailored for capturing temporal or shape-based similarities, resulted in compact clusters with reduced overlaps, especially in mid-to-high monetary segments.

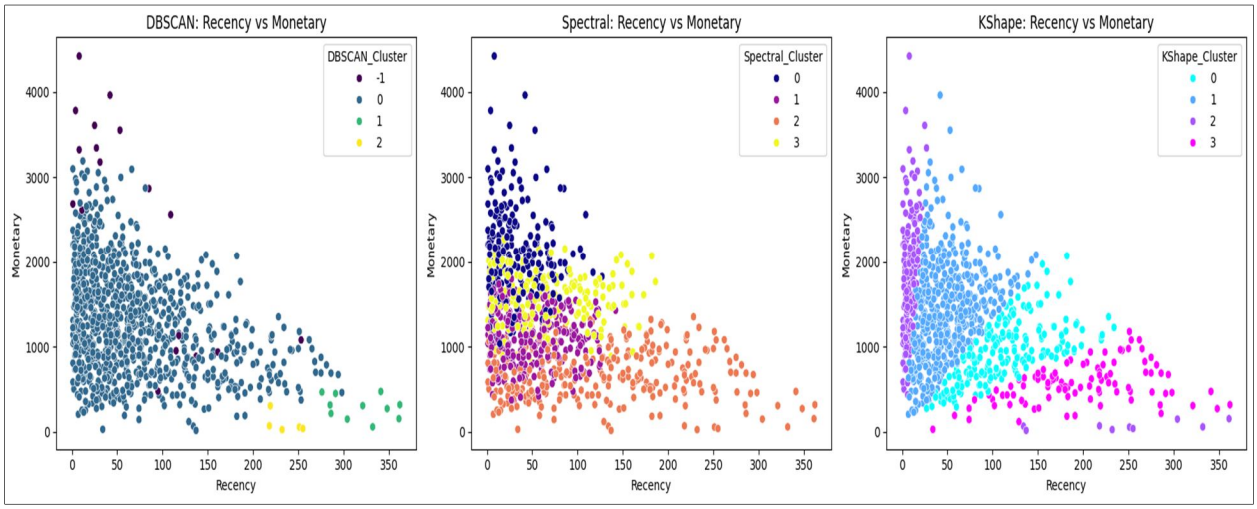


Figure 1: *2D Scatter Plots of Clustering Results on Recency and Monetary Dimensions*
(a) DBSCAN Clustering (b) Spectral Clustering (c) KShape Clustering

To gain deeper insight into the interrelations among Recency, Frequency, and Monetary value, 3D cluster visualizations were generated (Figure 2). DBSCAN (Figure 2a) again displayed strong density-based clusters but struggled to fully segment high-frequency spenders due to sparse data distribution. Spectral Clustering (Figure 2b) maintained clear separability in three dimensions, suggesting high compactness across all behavioral indicators. KShape (Figure 2c) showed effective segmentation, especially along the frequency axis, revealing its advantage in capturing temporal-spending cycles among customers.

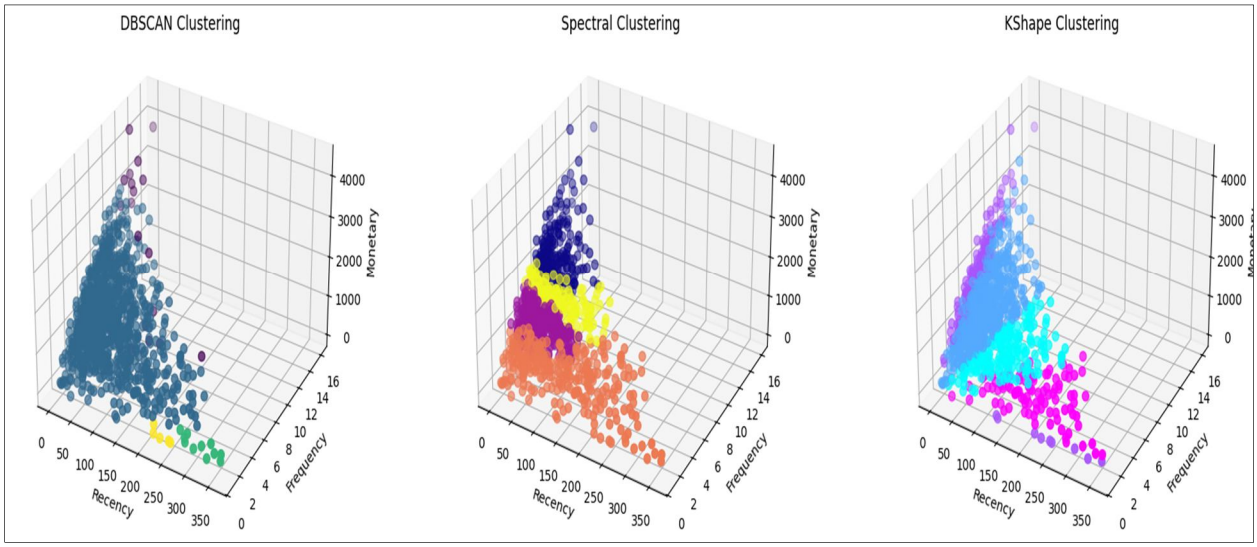


Figure 2: *3D Visualization of Clustering Results Using Recency, Frequency, and Monetary Features* (a) DBSCAN (b) Spectral Clustering (c) KShape Clustering

Silhouette analysis was performed to validate the internal quality of the clusters (Figure 3). DBSCAN (Figure 3a) exhibited a widespread in silhouette values, including several negative values caused by noise labels and poor separation among core points. Spectral Clustering (Figure 3b) achieved higher silhouette scores, confirming strong intra-cluster similarity and inter-cluster differentiation. KShape (Figure 3c) showed reasonably good silhouette values as well, with distinct peaks corresponding to the dense behavioral groupings it formed.

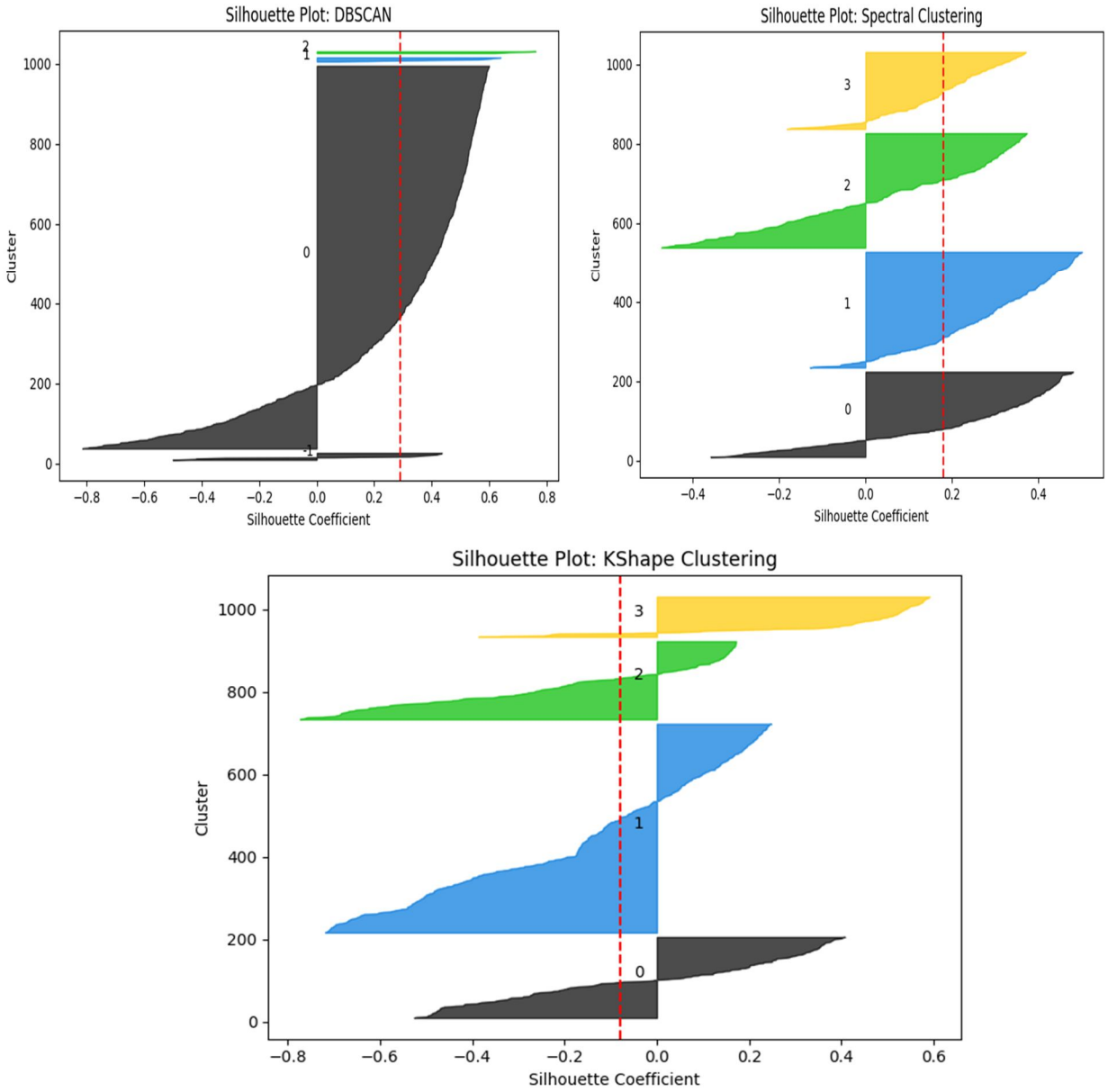


Figure 3: *Silhouette Analysis for Cluster Cohesion and Separation Across Models*
(a) DBSCAN (b) Spectral Clustering (c) KShape Clustering

Overall, the results indicate that KShape Clustering demonstrated superior performance in terms of alignment and interpretability. It achieved the highest accuracy (0.5091) and F1-score (0.6622) when evaluated against DBSCAN labels, highlighting its ability to replicate meaningful density-based patterns while incorporating temporal similarity. Additionally, its weighted precision remained high (0.9564), which is crucial for reducing misclassification in larger customer segments. While Spectral Clustering also performed well in terms of alignment, with a precision of 0.9629, its recall and F1-score were comparatively lower, indicating that fewer true clusters were captured despite confident predictions.

In unsupervised evaluation, Spectral Clustering achieved the lowest Davies-Bouldin Index (0.8015) and a Silhouette Score of 0.3417, suggesting strong intra-cluster cohesion and inter-cluster separation. DBSCAN, though achieving the highest Silhouette Score of 0.4316, had a poor Calinski-Harabasz Index (17.2542) and performed weakly in label alignment, reflecting limited interpretability despite compact clusters. KShape maintained a balance, with strong supervised scores and moderately competitive unsupervised metrics. Taken together, these findings suggest that KShape is the most robust model for this behavioral segmentation task, while Spectral Clustering remains a reliable alternative for scenarios prioritizing clean separation over temporal grouping.

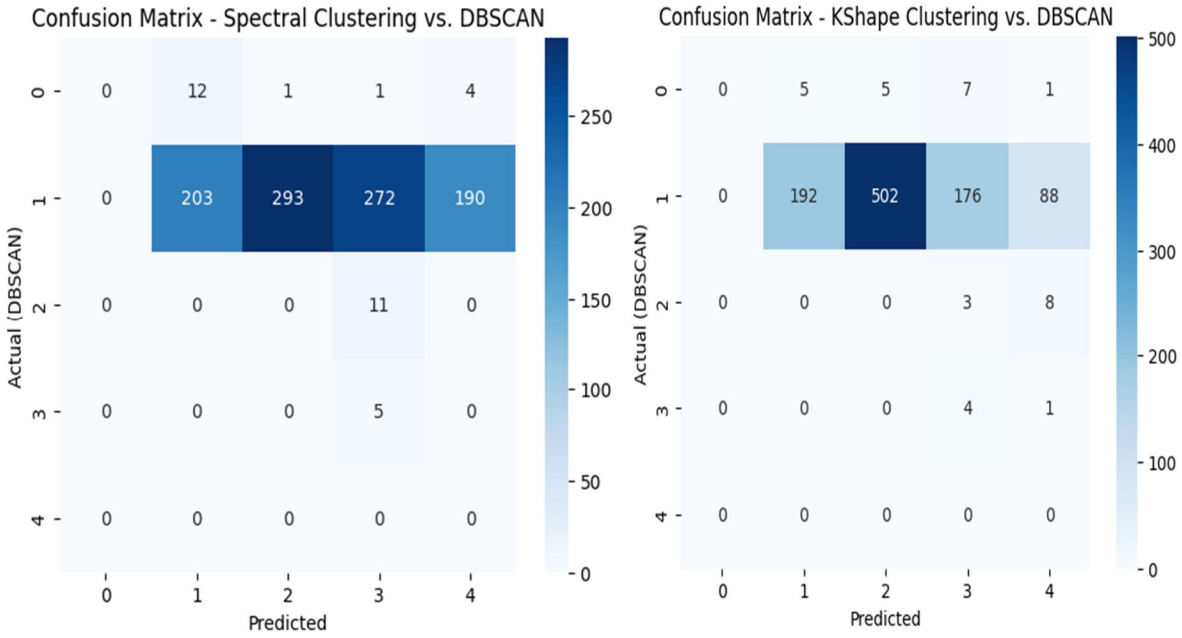


Figure 4: *Confusion Matrices Comparing Cluster Label Alignment with DBSCAN*
(a) Spectral vs. DBSCAN (b) KShape vs. DBSCAN

To quantitatively assess the clustering performance, three intrinsic evaluation metrics were employed: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. These metrics were computed for each of the six clustering algorithms applied on the standardized feature set.

Methods	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
KMeans	0.2578	1.2451	431.8579
Agglomerative	0.1876	1.3413	345.9340
GMM	0.1984	1.4315	287.2603
DBSCAN	0.4316	2.0641	17.2542
Spectral	0.3417	0.8015	92.0213

Table 1: Intrinsic clustering evaluation metrics across all models using Silhouette Score, Davies-Bouldin Index (lower is better), and Calinski-Harabasz Index (higher is better)

Among all models, DBSCAN achieved the highest Silhouette Score of 0.4316, indicating strong cohesion within clusters and clear separation between them. However, its Davies-Bouldin Index was also the highest (2.0641), suggesting some inter-cluster overlap, likely due to its noise handling. Spectral Clustering demonstrated a more balanced performance with a Silhouette Score of 0.3417 and the lowest Davies-Bouldin Index of 0.8015, reflecting high-quality clusters with minimal internal dispersion.

On the other hand, KMeans achieved the highest Calinski-Harabasz Index of 431.8579, highlighting well-separated and compact clusters under the assumption of convex boundaries. Both Agglomerative Clustering and GMM exhibited lower scores across all three metrics, suggesting less optimal segmentation performance in this specific context.

These results highlight the trade-offs between algorithms. While DBSCAN excels in internal cohesion, Spectral Clustering offers a balanced combination of structure and separation. KMeans remains competitive in scenarios favoring spherical cluster shapes. The evaluation confirms the complementary strengths of each algorithm, reinforcing the value of using diverse models for robust customer segmentation.

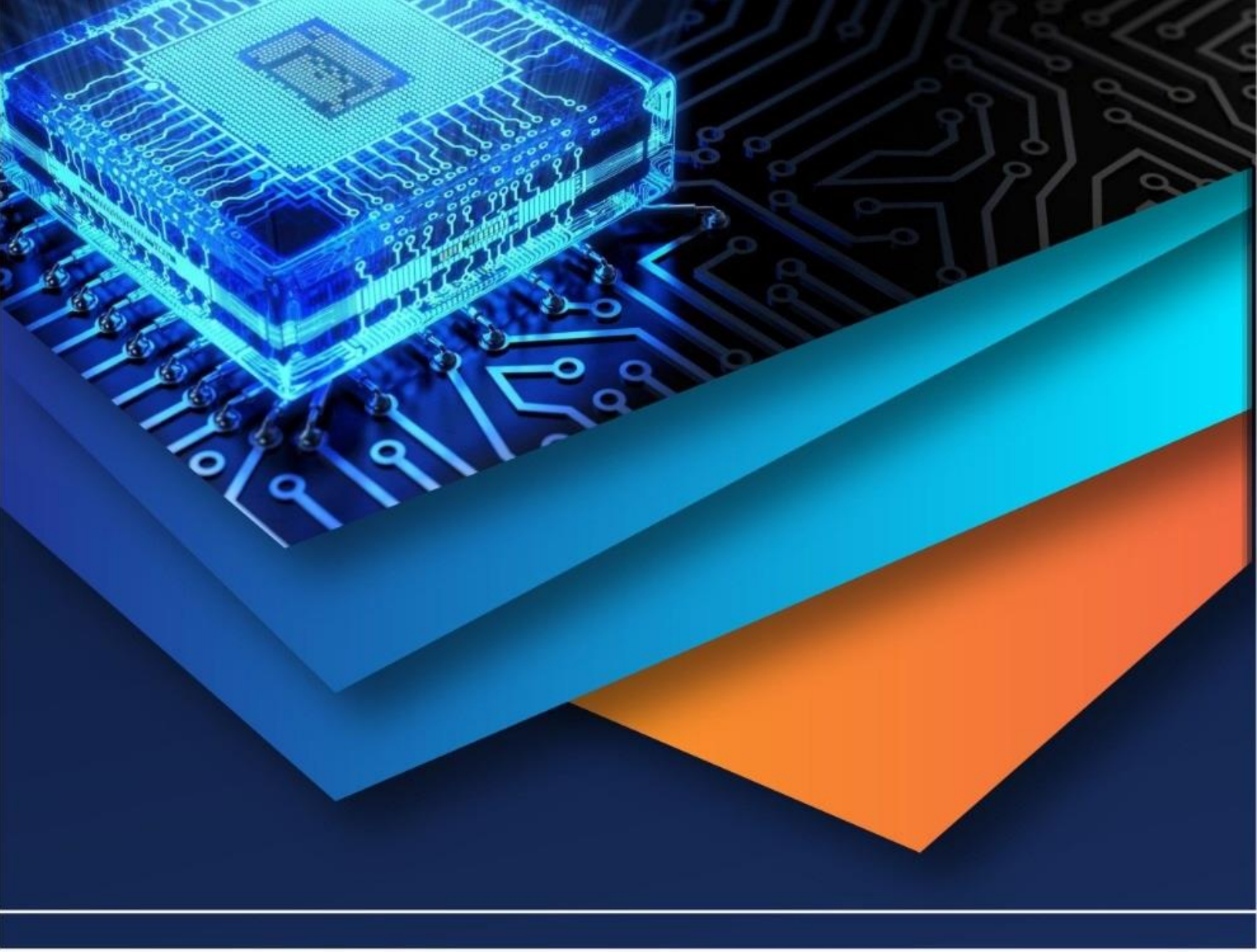
V. FINDINGS AND FUTURE SCOPE

The implementation of the extended LRFSM model enabled a more detailed understanding of customer behavior by incorporating five key behavioral metrics: Length, Recency, Frequency, Spend, and Monetary value. Clustering with six different algorithms revealed significant differences in customer group structures, with models like DBSCAN and Spectral Clustering showing strengths in identifying non-linear patterns and outliers, while KMeans and Mini-Batch KMeans offered more compact and stable clusters. Evaluation through both intrinsic metrics (Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz) and supervised metrics (Accuracy, Precision, Recall, F1-score) confirmed the robustness and effectiveness of the approach. Cluster profiling provided clear behavioral distinctions, supporting more personalized marketing strategies.

Looking ahead, future work may involve extending the framework using deep learning-based clustering methods such as Deep Embedded Clustering (DEC) for improved pattern recognition. Applying the model to real-world customer datasets with richer features—like purchase categories, session time, or profit—can enhance practical value. Additionally, exploring temporal behavior trends or developing hybrid and ensemble clustering approaches may offer even more robust segmentation. Automating and scaling the system for real-time segmentation in large-scale environments could further broaden its impact in practical e-commerce settings.

REFERENCES

[1] Al-Dubai, T. Al-Khalifa, S. Ahsan, R. F. Olanrewaju, and D. Yousif, “LRFS: Online Shoppers Behavior-Based Efficient Customer Segmentation Model,” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 14, no. 4, pp. 579–586, 2023.
[2] T. Lin, L. Sun, C. Guo, and Z. Zhang, “Adaptive Spectral Affinity Propagation Clustering,” Proceedings of the 2021 International Conference on Artificial Intelligence and Computer Engineering (ICAICE), 2021, pp. 478–483.
[3] H.-H. Zhao, X.-C. Luo, R. Ma, and X. Lu, “An Extended Regularized K-Means Clustering Approach for High-Dimensional Customer Segmentation With Correlated Variables,” IEEE Access, vol. 9, pp. 112508–112520, 2021.
[4] M. B. Hosseini and M. J. Tarokh, “Customer Segmentation Using CLV Elements,” International Journal of Engineering Research and Applications (IJERA), vol. 6, no. 2, pp. 68–74, 2016.
[5] J. Liao, A. Jantan, Y. Ruan, and C. Zhou, “Multi-Behavior RFM Model Based on Improved OM Neural Network Algorithm for Customer Segmentation,” IEEE Access, vol. 9, pp. 169389–169400, 2021.
[6] A. Naeem, A. Shabbir, N. U. Hassan, C. Yuen, A. Ahmad, and W. Tushar, “Understanding Customer Behavior in Multi-Tier Demand Response Management Program,” IEEE Access, vol. 8, pp. 170717–170733, 2020.
[7] J. W. T. M. de Kok et al., "Deep embedded clustering generalisability and adaptation for integrating mixed datatypes: two critical care cohorts," Scientific Reports, vol. 14, no. 1045, 2024, doi: 10.1038/s41598-024-51699-z.
[8] Y. G. Lee, C. W. Park, and S. J. Kang, "Deep Embedded Clustering Framework for Mixed Data," IEEE Access, vol. 10, pp. 1–12, 2022, doi: 10.1109/ACCESS.2022.3232372.
[9] X. D. Wang, R. C. Chen, F. Yan, Z. Q. Zeng, and C. Q. Hong, "Fast Adaptive K-Means Subspace Clustering for High-Dimensional Data," IEEE Access, vol. 7, pp. 42639–42652, 2019, doi: 10.1109/ACCESS.2019.2907043.
[10] N. T. Lee, H. C. Lee, J. Hsin, and S. H. Fang, "Prediction of Customer Behavior Changing via a Hybrid Approach," IEEE Open Journal of the Computer Society, vol. 1, pp. 1–12, 2023, doi: 10.1109/OJCS.2023.3336904.
[11] W. M. Wong and W. Su, "Segmenting Online Shoppers: A Combined Cluster and Logistic Regression Approach for Forecasting Purchase Behavior," IEEE Access, vol. XX, pp. 1–10, 2025, doi: 10.1109/ACCESS.2025.3565897.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY