Proceedings of the International Conference on AI-Driven Smart Systems and Ubiquitous Computing (ICAUC-2026)
DVD Part Number: CFP266AK-DVD; ISBN: 979-8-3315-5850-5

Customer Behavior Analytics Using Manifold Learning and Ensemble Clustering

Ms R.Anitha
Research Scholar,
Department of Computer Science
Vels Institute of Science, Technology
Advanced Studies(VISTAS)
E-mail:anithashivaguru@gmail.com

Dr. Y.Kalpana
Professor,
Department of BCA & IT
Vels Institute of Science, Technology
Advanced Studies(VISTAS)
E-mail: kalpana.scs@vistas.ac.in

***Abstract*—Customer segmentation plays a crucial role in analysing purchasing patterns of customers in large scale retail industry. However, transaction data are usually high-dimensional and noisy which limits the effectiveness of conventional clustering techniques. This paper presents an integrated framework that combines manifold learning and ensemble clustering to make improved and robust customer behavior analysis. Initially RFM and other insightful purchase behavior features are extracted from the transaction record of online retail dataset. UMAP is then applied to reduce the dimension and also preserves the intrinsic data structure. HDBSCAN is employed to identify high density clusters and detect noise. To enhance quality of cluster, MiniBatch K-Means is applied in core clusters and a GMM assigns noisy customers based on probabilistic similarity. The experimental results on an online retail dataset demonstrate the robust and interpretable customer segments. The proposed framework improves clustering stability and also supports effective data-driven customer analytics and targeted marketing strategies.**

Keywords:- Customer segmentation, Clustering Algorithm, Machine Learning, Density-Based Clustering, Centroid-Based Clustering, Data Analytics, UMAP, Ensemble Clustering.

I. INTRODUCTION

The rapid growth of e-commerce and digital retail platforms has led to the large volumes of transactional data describing customer purchasing behaviour. Analysing this data effectively is essential for understanding consumer preferences, improving customer engagement, and supporting applications such as personalized marketing, recommendation systems, and customer relationship management. Customer segmentation has become a core task in retail market, as it allows business to group customers based on shared behavioural characteristics and to develop targeted strategies [1], [2].

Customer segmentation is challenging in the existing retail environments. Transactional datasets are typically high-dimensional data with sparse, and noisy, with customers demonstrating irregular purchase frequencies, varying spending levels, and nonlinear behavioural relationships. Traditional clustering algorithm such as K-Means, hierarchical clustering are still widely used because of their simplicity. However, they depend on linear separability and uniform cluster structure, which rarely hold in practice. Such methods often produce unstable clusters when applied to complex, real world data [3]. Linear dimensionality reduction such as Principal Component Analysis (PCA) can further simplify the data but fail to preserve the essential nonlinear structure present in customer behaviour patterns [4].

To address these limitations, density-based clustering approach the extension of DBSCAN that is HDBSCAN have been used, as they are capable of identifying clusters with arbitrary shapes and detecting noise [5], [6]. This method improve flexibility, frequently classify a large proportion of customers as noise and are sensitive to parameter selection, limiting their direct applicability in business scenarios when complete and interpretable segmentation is required. Recent advances in manifold learning, particularly Uniform Manifold Approximation and Projection UMAP, have demonstrated strong capability in capturing nonlinear data structures while preserving both local and global relationships [7]. However, UMAP is often used only as a visualization tool and is rarely integrated into a complete clustering framework for customer behaviour analysis.

Driven by these challenges, this research proposes an integrated manifold learning and ensemble clustering framework for high-resolution customer behaviour analysis. The framework combines UMAP for nonlinear dimensionality reduction with HDBSCAN for density-based cluster discovery, followed by boundary refinement using MiniBatch K-Means and probabilistic customer assignment through Gaussian Mixture Models. By integrating complementary techniques, the proposed approach captures detailed behavioural patterns, improves cluster stability, resolves noise related issues, and ensures complete customer coverage. Experimental evaluation on a online retail dataset demonstrates that the proposed framework produces more stable, interpretable, and

business relevant customer segments compared to traditional segmentation techniques.

II. LITERATURE REVIEW

A. *Foundation: RFM and Clustering Limits*

The RFM (Recency, Frequency, Monetary) framework is used in customer segmentation to turn raw transaction logs into behavioral features. Classic methods like K-Means are important, but they often have trouble with the "messiness" of real-world data. As Lewaa [9] points out, K-Means assumes clusters are spherical, which rarely happens in retail. Recent work by Asana et al. [10] tried to fix this using Fuzzy C-Means to allow for "overlapping" customers, but even these models are sensitive to outliers. The consensus in recent studies [11] is that traditional algorithms are too rigid for high-dimensional data, creating a need for the more flexible manifold-based approaches used in this study.

B. *Dimensionality Reduction via UMAP*

The leap from linear methods like PCA to non-linear manifold learning has changed how we handle customer "features." Allaoui et al. [12] demonstrated that UMAP (Uniform Manifold Approximation and Projection) is significantly better at keeping the global structure of a dataset intact while reducing noise. Unlike older methods, UMAP "unfolds" the data, making it easier for density-based algorithms to see the natural groupings. Oide and Sugita [13] confirmed that this projection is a critical first step because it effectively "cleans" the high-dimensional space before any clustering begins.

C. *The Challenge of Noise in HDBSCAN*

Density-based clustering, specifically HDBSCAN, is favoured in recent research because it doesn't force you to pick the number of clusters (k) in advance. Wellén and Solbu [14] used this for banking data and found segments that distance-based models missed. However, there is a catch: HDBSCAN is very strict. It often labels a huge chunk of the customer base as "noise" (outliers). Blachowicz et al. [15] argue that while this is mathematically sound, it's a problem for businesses that can't just ignore 15-20% of their customers. This gap—where robust density detection meets the practical need for 100% coverage—is exactly where this research fits in.

D. Hybrid Approaches for Total Coverage

To solve the noise problem, researchers are starting to combine models. The latest trend involves "probabilistic recovery." Instead of throwing away the noise points, we can use Gaussian Mixture Models (GMM) to give them a second look. Guo et al. [16] suggest that since GMM works on probability, it can assign these "uncertain" points to the most likely group rather than leaving them unclassified. By using Mini-Batch K-Means for the core groups and GMM for the noise, we get the best of both worlds: the speed and precision of density clustering with the full coverage of probabilistic models [17], [18].

Most existing UMAP- HDBSCAN methods only look for clusters and avoid noise points which makes them less useful in real life. This proposed method fixes this gap by adding an ensemble refinement strategy to the basic UMAP-HDBSCAN method. The clusters after refining the noise points makes the customer segments more stable and complete for customer analytics.

III. METHODOLOGY

This methodology is used to address challenges associated with high dimensionality, nonlinear data structure, and noise which is present in online retail datasets. The whole process includes cleaning up the data, extracting features to analyze customer purchase behavior, reducing the number of dimensions using UMAP, clustering based on density with HDBSCAN, improving the cluster by ensemble with MiniBatch K-Means and Gaussian Mixture Models, and finally evaluating the segments. The complete workflow is illustrated in Fig. 1 and described in the following subsections.

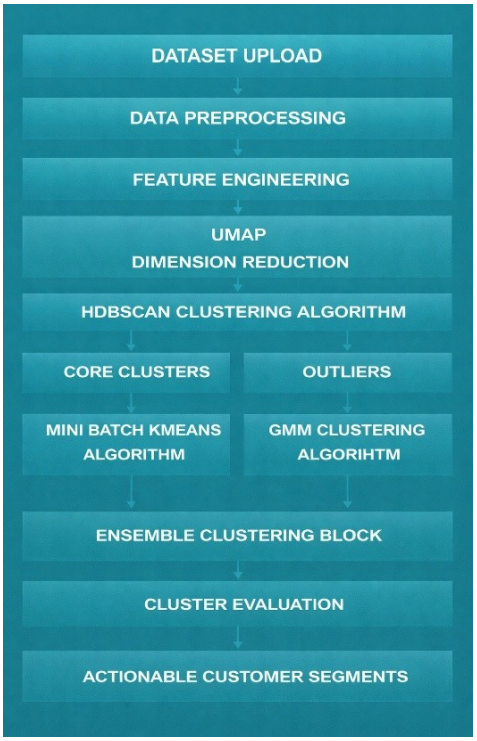

Fig 1. Workflow of Proposed Model

A. Dataset Description and Upload

This analysis is conducted using an online retail transaction dataset consisting of customer purchase records, including invoice dates, product quantities, unit prices, and customer identifiers. This online retail dataset is uploaded and organized to apply it as the raw input for subsequent preprocessing and analysis.

B. Data Preprocessing

Data preprocessing is performed to make the data quality and reliability for analysis. The process involved in this stage are removal of duplicate records, handling of missing customer identifiers, elimination of canceled transactions, standardize the transaction timestamps and cleaning of invalid or negative quantities. Then the processed data is aggregated at the customer level to enable robust behavioral analysis.

C. Feature Engineering

The features used in this study are chosen to describe customer purchasing behaviour in a simple and meaningful way. Recency state how recently a customer has made a purchase, which helps to identify active and inactive customers. Frequency shows how often a customer buys, while Monetary reflects the total amount spent over time. Together, these features provide a basic but effective understanding of customer engagement and value.

TABLE I. EXTRACTED FEATURES

Extracted Feature	Description
Average Purchase Interval	Average of differences between sorted unique purchase dates.
Product Diversity	Count of unique products bought be each customer
Average Quantity	Average quantity of items purchased per transaction
Total Quantity	Sum of all quantities purchased by a customer
Average Monetary per Purchase	Average transaction value per purchase
Days since Last Purchase	Difference in days between last invoice date per customer and overall max invoice date

To acquire behaviour in more detail, features to be extend beyond the traditional RFM model. Average Purchase Interval describes how regularly customers make purchases, and also used to differentiate frequent shoppers from occasional ones. Product Diversity represents the variety of products purchased. Average Quantity and Total Quantity calculate the volume of purchase, while Average Monetary Value shows the average spend per transaction. Combined, these features give a clear and more realistic customer buying patterns, supports more accurate and interpretable segmentation for data analytics.

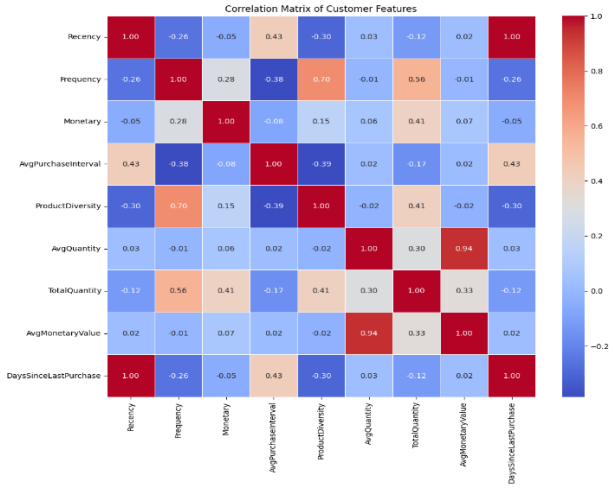

Fig 2. Correlation Matrix of Initial Feature Set

In the first set of features, there is a strong overlap between Recency and DaysSinceLastPurchase, which means that both features give the identical information. In the same way, AvgQuantity and AvgMonetaryValue have a very strong positive correlation, which shows that quantity based and spending based measures share some information. The existence of such robust correlations validates multicollinearity within the initial feature space.

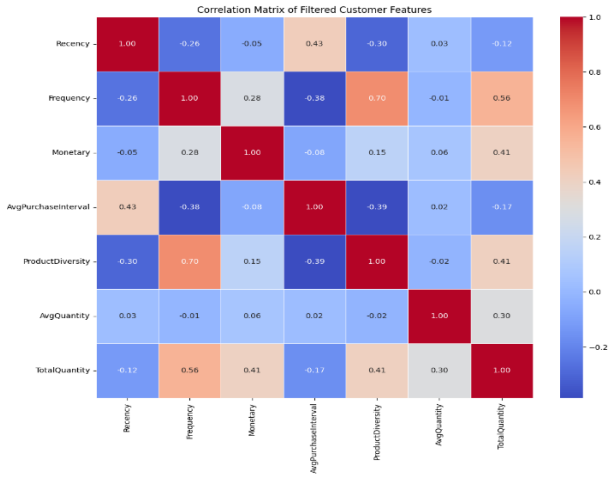

Fig 3. Correlation Matrix of Filtered Feature Set

The filtered correlation matrix displays a more equitable correlation structure subsequent to the elimination of highly correlated features. There are still important behavioral relationships, but high linear dependencies are lessened. For instance, Frequency is still positively correlated with ProductDiversity and TotalQuantity. This means that people are still buying things in the same way. AvgPurchaseInterval is still negatively correlated with Frequency. This means that customers who don't buy things as often have longer gaps between purchases. The filtered feature set shows that features are less similar and more independent from each other. This set was used as input for UMAP-based dimensionality reduction and then for clustering with multiple groups.

D. *Dimensionality Reduction using UMAP*

To reduce the high dimensionality and handle nonlinear structure of customer behavior dataset, Uniform Manifold Approximation and Projection (UMAP) is proposed. UMAP projects the reduced feature space into a low-dimensional manifold by preserving local neighborhood relationships and also global structure. This provides a suitable representation for density-based clustering.

E. *UMAP Embedded HDBSCAN Clustering*

UMAP embedded HDBSCAN algorithm is performed to produce initial clusters. It identifies clusters of varying density without initiating the number of clusters. The resulting clusters reveal distinct segments based on purchase behavior features.

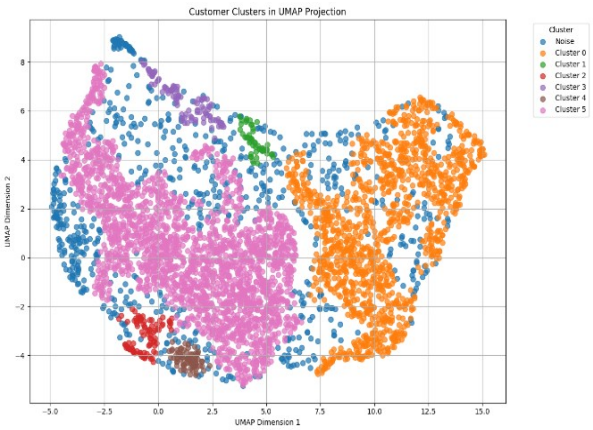

Fig 4. UMAP Embedded HDBSCAN Clustering

Fig 4 The plot picturises the cluster formation after applying HDBSCAN on the UMAP produced dataframe. Each point in the graph represents a distinct customer, each color indicates different cluster. The separated dense region by forming clusters is observed. This indicates that UMAP captures the structure of customer purchase behavior effectively.

Meaningful customer segmentation is done by HDBSCAN algorithm and also marks the scattered data points as noise. We can notice that varying purchase patterns among customers by the presence of compact clusters and isolated noise points in the plot. This visualization clearly states that combined use of UMAP and HDBSCAN provides a clear and interpretable initial segmentation of customers.

F. *Mini Batch K-Means Clustering*

Fig. 5. depicts that customer segmentation is obtained by applying MiniBatch K-Means clustering algorithm to the core clusters produced by HDBSCAN algorithm on the UMAP embedding.

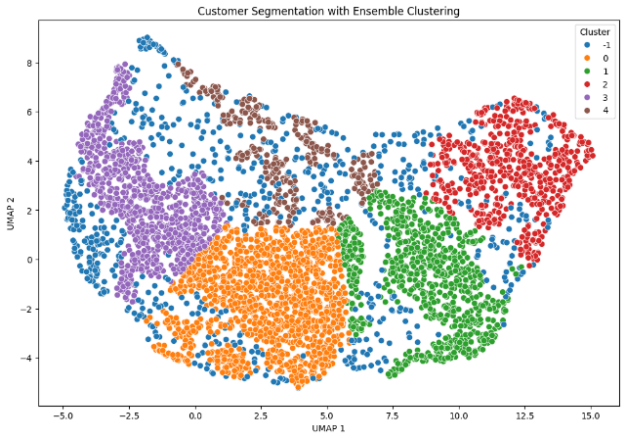

Fig 5. MiniBatch K-Means Clustering

Compactness of clusters can be improved by applying MiniBatch K-Means to refine the dense region detected by HDBSCAN. The resulted clusters appear more coherent and well separated. It also preserves the original structure learned from UMAP. This refinement helps to reduce cluster overlap and enhances scalability for large datasets.

G. *Noise Refinement using GMM*

After the forming compact clusters using MiniBatch K-Means by refining the core clusters, Gaussian Mixture Model (GMM) is applied to the remaining noise points identified by HDBSCAN. GMM refine the noise part identified by HDBSCAN to analyse all the customers purchase behaviour. GMM allocates these noise samples to the most probable clusters based on their probability in the feature space. This probabilistic assignment reduce the scattering, reduces number of unclassified customers and improves cluster completeness. As a result, uncertain customers are meaningfully integrated to form new clusters leading to more stable and interpretable customer segments.

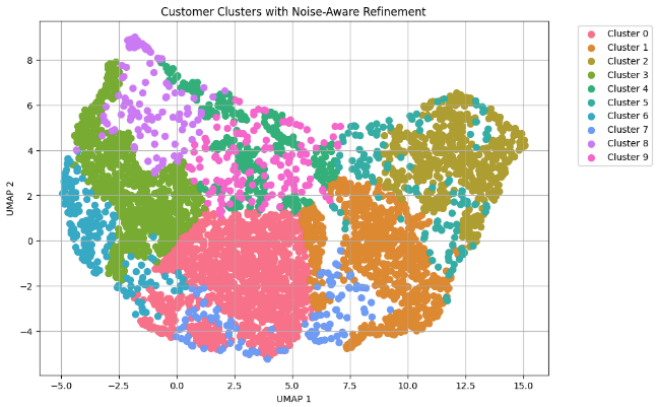

Fig 6. GMM Clustering

Fig. 6. shows that refined clusters are compact and well separated. It shows that customers previously labelled as noise are assigned probabilistically to the suitable clusters. The increase in the number of well-defined clusters shows behavioral differentiation among customers. Each cluster is a homogeneous group with similar purchasing patterns. In this refinement, the valuable customers are not discarded as noise. The final segmentation provides a balance between density-based discovery and centroid-based refinement which leads to actionable and reliable customer segments.

H. *Cluster Evaluation*

The quality of customer segments is evaluated using validation metrics like Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Index. Cluster stability is evaluated using measures such as Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). These metrics can be collectively assessing compactness, separation, and robustness of the ensemble clustering results.

Average membership probability for clustered points: 0.9804
Silhouette Score: 0.3310
Davies-Bouldin Index: 0.9529
Calinski-Harabasz Index: 3204.5183
Normalized Mutual Information: 0.71
Adjusted Rand Index: 0.64

I. *Generation of Customer Segments*

The mean values of customer purchase behaviour features across the clusters which differentiate the purchasing patterns clearly. Cluster 3 represents most valuable customer group with the lowest recency, high frequency, extremely high monetary, high product diversity and also high total quantity. This purchasing behavior of customers in this cluster reflects strong loyalty and sustained engagement.

Clusters 8 represents high monetary and total quantity but high recency and less frequency compared to cluster 3. This indicates the customers in this cluster are high-spending with large quantities but less frequently.

Clusters 6 and 0 are relatively low recency, moderate frequency along with balance monetary and product diversity. This indicates that the customers in these clusters are stable and consistent. Cluster 4, 7 and 9 represents moderate recency, moderate monetary and limited product diversity.

In contrast, Clusters 2 and 5 exhibits very high recency, low frequency, low monetary and minimal product diversity. This indicates that the customers in these clusters are likely to be inactive or churn-prone. Their average purchase intervals and lower quantities further confirm the weak engagement of customers. Overall, the segmentation captures not only customer value but also purchasing regularity, volume, and diversity. This provide a rich and more actionable understanding of customer purchase behavior.

IV. CONCLUSION

This work demonstrates an integrated customer segmentation that combines manifold learning and ensemble clustering techniques to analyse high-dimensional retail industry dataset effectively. Beyond traditional RFM, the extended purchase behavioral features are incorporated to capture customer activity like spending patterns, purchase regularity, volume and product diversity in a unified manner. The influence of additional features also enhances the conventional RFM feature set. The results indicate that new attributes yield supplementary behavioral insights that improves cluster differentiation and unity. This analysis shows that detailed features make customer segments more meaningful and stable that makes the proposed framework even more effective.

The application of UMAP enables meaningful dimensionality reduction. The HDBSCAN clustering technique identifies the natural customer groups and also isolates irregular purchase behavior.

The compactness, stability and completeness of clusters improved more by refining the clusters by applying MiniBatch K-Means and GMM clustering algorithms. The Evaluation metrics and cluster profiling demonstrate clear separation between high-value, moderate and inactive customer segments. Overall, this proposed framework produces interpretable and actionable customer segments. It is suitable for large-scale customer analytics, data-driven decision making and also targeted marketing in online retail environments.

REFERENCES

[1] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, “Application of data mining techniques in customer relationship management: A review and classification,” Expert Systems with Applications, vol. 36, no. 2, pp. 2592–2602, Mar. 2009, Doi: 10.1016/j.eswa.2008.02.021.

[2] Michel Wedel and Wagner A. Kamakura, Market Segmentation: Conceptual and Methodological Foundations, 2nd ed. Boston, MA, USA: Kluwer Academic Publishers, 2000, doi: 10.1007/978-1-4615-4651-1.

[3] Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2011.

[4] Anil K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, Jun. 2010, doi: 10.1016/j.patrec.2009.09.011.

[5] Harold Hotelling, “Analysis of a complex of statistical variables into principal components,” Journal of Educational Psychology, vol. 24, no. 6, pp. 417–441, 1933, doi: 10.1037/h0071325.

[6] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD), Portland, OR, USA, 1996, pp. 226–231.

[7] Leland McInnes, John Healy, and Steve Astels, “HDBSCAN: Hierarchical density based clustering,” Journal of Open Source Software, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.

[8] Leland McInnes, John Healy, and James Melville, “UMAP: Uniform manifold approximation and projection for dimension reduction,” arXiv preprint arXiv:1802.03426, 2018.

[9] I. Lewaa, "Customer Segmentation Using Machine Learning Model: An Application of RFM Analysis," Journal of Business Analytics, vol. 6, no. 2, pp. 112–124, May 2023. [Online]. Available: https://doi.org/10.1080/2573234X.2023.2185412

[10] I. M. D. P. Asana, I. G. I. Sudipa, and K. J. Atmaja, "Customer Segmentation Using RFM Model and Fuzzy C-Means at PT SNS 21 Bali," International Journal of Intelligent Systems and Applications in Engineering, vol. 13, no. 1, pp. 45–52, Jan. 2025.

[11] M. I. Alhassan and R. Asare, "Enhancing Customer Segmentation Using Hybrid RFM-Fuzzy Clustering Approach in the Retail Industry," International Journal of Data Science and Advanced Analytics, vol. 5, no. 3, pp. 210–225, 2023.

[12] M. Allaoui, M. L. Kherfi, and A. Cheriet, "Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique," IEEE Access, vol. 10, pp. 56123–56135, May 2022. [Online]. doi: 10.1109/ACCESS.2022.3177309.

[13] S. Oide and M. Sugita, "Dimensionality Reduction using UMAP for Clustering of High-Dimensional Data," Computational Statistics & Data Analysis, vol. 172, p. 107489, Aug. 2022. [Online]. doi: 10.1016/j.csda.2022.107489.

[14] W. Wellén and D. Solbu, "Segmenting Bank Customers to Explore Product Engagement Opportunities: A Comparative Study Combining Clustering and Predictive Modeling," Master’s thesis, Dept. Business Administration, Umeå Univ., Umeå, Sweden, 2025.

[15] A. Blachowicz, P. K. Jonecko, and M. G. Kozielski, "Improving the Performance of HDBSCAN on Short Text Clustering by Using Word Embedding and UMAP," Expert Systems with Applications, vol. 240, p. 122481, Apr. 2024. [Online]. doi: 10.1016/j.eswa.2023.122481.

[16] Q. Guo, X. Liu, and J. Zhang, "UMAP-Based Clustering Split for Rigorous Evaluation of AI Models in Chemical Informatics," Cambridge Open Engage, Jun. 2024. [Online]. doi: 10.33774/chemrxiv-2024-p88kj.

[17] R. Zhang, S. Wang, and T. L. J. Howard, "Dynamic Customer Segmentation: From RFM to Deep Clustering and Manifold Learning in Digital Markets," Journal of Marketing Analytics, vol. 13, no. 1, pp. 12–29, Mar. 2025. [Online]. doi: 10.1057/s41270-024-00305-w.

[18] A. Ahmad and S. Khan, "A Survey of Hybrid Clustering: Challenges and Opportunities," IEEE Transactions on Knowledge and Data Engineering, vol. 36, no. 4, pp. 1540–1558, Apr. 2024. [Online]. doi: 10.1109/TKDE.2023.3289012.

[19] Wang, Chenguang. "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach." Information Processing & Management 59.6 (2022): 103085.

[20] Zong, Yi and Enze Pan. “A SOM-Based Customer Stratification Model.” Wireless Communications and Mobile Computing 2022.

[21] Shirole, Rahul, Laxmiputra Salokhe, and Saraswati Jadhav. "Customer segmentation using rfm model and k-means clustering." Int. J. Sci. Res. Sci. Technol 8.3 (2021): 591-597.