VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY, NAMBUR, AP, INDIA
(An Autonomous Institution permanently affiliated to JNTUK - Approved by AICTE New Delhi - Accredited by NBA - NAAC with 'A' Grade - All eligible branches are Accredited by NBA - ISO9001:2015 Certified)
Department of Computer Science & Engineering
Proceedings of International Conference on Advanced Computing Technologies (ICACT-2025), 3rd April 2025 & 4th April 2025, ISBN: 978-81-977391-1-8

Customer Segmentation using DEC

K. Mohan Krishna[1], N. Manasa[2], P. Trisha[3], P. Sai Ganesh[4], P. Bhavana[5]
[1]Associate Professor, Department of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, (Autonomous), Guntur, Andhra Pradesh, India, [1]mohankrishnakotha@gmail.com
[2,3,4,5] Students, Department of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, (Autonomous), Guntur, Andhra Pradesh, India.
narramanasa61@gmail.com, pothinatrisha@gmail.com, psaiganesh7728@gmail.com, pbhavana2004@gmail.com

Abstract- Customer segmentation is a strategic marketing approach that divides a company's customer base into distinct groups based on shared characteristics, behaviors, needs, or preferences. This practice allows businesses to tailor their marketing efforts, product development, and customer service approaches to address the specific requirements of different customer groups more effectively.
Deep Embedded Clustering (DEC) is applied for customer segmentation in contemporary marketing environments. It represents an advanced approach that combines deep neural networks with clustering algorithms to discover latent patterns in high-dimensional customer data that traditional segmentation methods often fail to identify. Our methodology employs an autoencoder architecture to learn low-dimensional representations of customer features, followed by a specialized clustering algorithm that iteratively refines both the feature representations and cluster assignments. This approach enables the identification of complex, non-linear relationships among customer attributes including transactional history, digital engagement metrics, psychographic variables, and cross-channel behaviours.

Keywords: Forecasting, RFM Analysis, K-Means, Hierarchical Clustering, DBSCAN, DEC.

I. INTRODUCTION

With the fast rise of e-commerce websites, companies are gathering humongous data about customers daily. This data contains rich insights into customer preference, shopping, and buying patterns. It is crucial to comprehend these behaviours for developing customized marketing strategies and enhancing customer satisfaction. For this, companies tend to segment customers by their shared features. Conventional segmentation techniques such as K- Means and DBSCAN are widely employed for the same. But they are limited in identifying subtle patterns in intricate data, particularly when data is high-dimensional or non- linear.
New developments in deep learning have unlocked new potential in customer segmentation. One of those methods is Deep Embedded Clustering (DEC), which unites feature extraction and clustering into one process. DEC employs autoencoders — a form of neural network — to compress customer information into lower-dimensional, more representative forms, while at the same time clustering similar customers. This method works particularly well in uncovering latent patterns that other methods may overlook. In this project, we suggest a customer segmentation model based on DEC applied to e-commerce datasets. The model adopts Recency, Frequency, and Monetary (RFM) values to signify customer behaviour. Our approach entails data preparation, training an autoencoder to decrease its complexity, and clustering customers according to the DEC framework. We contrast the results of DEC with classic approaches in order to show its better performance in determining more significant customer segments.
The rest of this paper follows the following order: Section II gives an overview of related literature in customer clustering and segmentation. Section III details the proposed method, which encompasses data preprocessing as well as the DEC framework. Section IV summarizes the experimental outcomes and comparative results. Section V concludes the paper and proposes possible future research directions.

II. RELATED WORKS

Customer datasets are implemented and converted effectively to a CSV file to support mathematical and statistical analysis. Meditation is about dividing customers into groups, depending on behaviour and vans for expenses. This involves analysing analysis that he recently acted, how many times they buy, and how much they spend. To make the insight easier to understand, the results will be shown using 3D visualization. This visualization will provide a clear picture of customer groups. A deep built -in grouping (Dec) algorithm is used for this purpose, which combines deep learning with clustering techniques. Sci-kit, pandas and food plot lib are some libraries used on top of the operations described

A. Customer Classification

In today's competitive business environment, organizations

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY, NAMBUR, AP, INDIA
(An Autonomous Institution permanently affiliated to JNTUK - Approved by AICTE New Delhi - Accredited by NBA - NAAC with 'A' Grade - All eligible branches are Accredited by NBA - ISO9001:2015 Certified)
Department of Computer Science & Engineering
Proceedings of International Conference on Advanced Computing Technologies (ICACT-2025), 3rd April 2025 & 4th April 2025, ISBN: 978-81-977391-1-8

face increasing pressure to meet their customers' diverse needs and preferences, attract new people and increase their overall business performance. However, catering to each individual customer is a challenging task due to their priorities, needs, demographics, taste and variation in other factors. Treating all customers equally is often an effective business strategy. To address this challenge, business customers adopt the concept of division or market division, including dividing consumers into separate groups or segments. Each subgroup is made up of customers who share similar characteristics or display comparable behaviour in the market. Customer partitions help businesses better understand and target their audiences, allowing more tailored and effective strategies to meet specific customer needs.

A. Data Repository

The data collection involves systematically collecting and measuring information to track targeted changes within a given system. This process helps to answer relevant questions and evaluate the results effectively. It is an important component of research in various subjects including physics, social science, humanities and business. The primary goal of data collection is to achieve reliable and quality evidence that facilitates accurate analysis and leads to clear, fair answers of research questions. For this project, data was obtained from Kaggle repository.

B. Data Preprocessing

Preprocessing steps were performed to enhance data quality and reliability:

Handling Missing Values: Missing entries were addressed using mean/mode imputation. Normalization: Continuous variables were scaled using Min-Max normalization:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Feature Engineering: Derived attributes such as customer lifetime value (CLV) and churn probability.

Outlier Detection: Utilized the IQR method:

$$IQR = Q_3 - Q_1,\ X_{outlier} = X < Q_1 - 1.5 \times IQR \text{ or } X > Q_3 + 1.5 \times IQR$$

The formula for calculating a z-score is $z = (x-\mu)/\sigma$, where x is the raw score, μ is the population mean, and σ is the population standard deviation.

Exploratory Data Analysis (EDA) is conducted to analyse patterns and trends, including calculating the number of unique products, examining product quantities, and understanding customer behaviour. Visualizations like histograms aid in uncovering relationships and distributions, ensuring the dataset is well-prepared for analysis.

C. Feature Engineering

Feature engineering involves creating meaningful features that improve analysis and model performance. RFM analysis is considered as a most popular feature selection method, especially in the context of customer segmentation and behaviour analysis.

It involves transforming raw transactional data into three meaningful features: Recency, Frequency, and Monetary value. Recency measures how recently a customer made a purchase, calculated as the difference between the most recent transaction date and the current date, indicating customer engagement and activity levels. Frequency captures the number of transactions a customer has made, reflecting loyalty and satisfaction. Monetary value represents the total amount a customer has spent, highlighting high-value customers.

D. Data Clustering

The clustering is the technique of organizing data in groups or groups based on shared characteristics or equality within the dataset. Depending on specific conditions and requirements of data, various algorithms can be used for clustering. However, there is no single clustering algorithm that works universally for all datasets. Therefore, it is important to select the most suitable clustering method. In this work, we implemented four clustering algorithms using the Python Scikit-Library.

III. METHODOLOGY

Segmentation Techniques—Several clustering methods were evaluated:

K-Means Clustering: K-Means was used for customer segmentation based on RFM values. The cost function minimized is:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n} \| x_j - c_i \|^2$$

where ci represents the centroid of cluster i. The optimal number of clusters was determined using the Elbow Method and Silhouette Score.

DBSCAN: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was applied to identify core and noise points. The algorithm defines core points based on a minimum number of neighbours within a radiusƐ.

Hierarchical Clustering: Hierarchical clustering was employed to build a dendrogram- based customer segmentation. The linkage function is defined as:

$$D(a, b) = \min \|x_i - x_j\|,\ x_i \in A,\ x_j \in B$$

Where D (a, b) is the distance between clusters A and B.

RFM Analysis: Customers were classified into distinct groups using Recency, Frequency, and Monetary (RFM) scores.

$$RFM = w_1R + w_2F + w_3M$$

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY, NAMBUR, AP, INDIA
(An Autonomous Institution permanently affiliated to JNTUK - Approved by AICTE New Delhi - Accredited by NBA - NAAC with 'A' Grade - All eligible branches are Accredited by NBA - ISO9001:2015 Certified)
Department of Computer Science & Engineering
Proceedings of International Conference on Advanced Computing Technologies (ICACT-2025), 3rd April 2025 & 4th April 2025, ISBN: 978-81-977391-1-8

where w1, w2, w3 are the assigned weights for recency, frequency and monetary values.

Deep Embedded Clustering:

DEC integrates feature learning and clustering into a single framework using autoencoders. The loss function for clustering refinement is defined as:

$$L = KL(P||Q) = \sum p_{ij} \log \frac{p_{ij}}{q_{ji}}$$

where p_{ij} is the probability of assigning point x_i to cluster j and q_{ij} is the soft cluster assignment obtained via a student's distribution. Clustering effectiveness was assessed using: Silhouette Score:

$$S = \frac{b - a}{\max(a, b)}$$

where a is the mean intra-cluster distance, and b is the mean nearest-cluster distance.

Elbow Method: Determined the optimal number of clusters based on the within-cluster sum of squares.

Davies-Bouldin Index: Evaluated inter-cluster similarity.

Domain Validation: Expert validation ensured practical interpretability

Implementation and Insights:

The segmented data provided actionable insights for targeted marketing strategies. Customer groups were profiled based on spending behavior, engagement, and churn probability. Visual analysis, including RFM plots, dendrograms, and silhouette plots, was conducted to validate segmentation effectiveness.

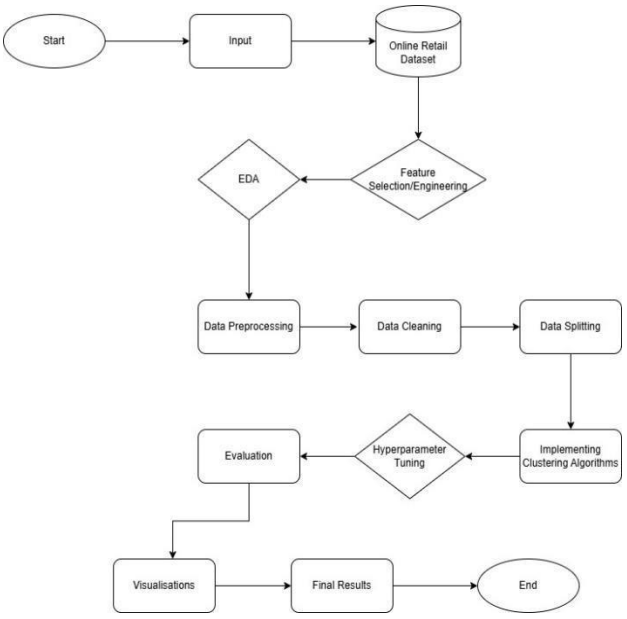

Figure 1. System Architecture Diagram

IV. RESULTS AND DISCUSSIONS

The proposed method, combining RFM-based customer segmentation with Deep Embedded Clustering (DEC), was evaluated using multiple performance metrics. The dataset was preprocessed by handling missing values, normalizing features, and selecting relevant attributes essential for clustering. Various clustering techniques such as K-Means, Hierarchical Clustering, DBSCAN, and Deep Autoencoder-based Clustering were applied, and their results were analyzed.

The clustering results, as visualized in Figures 1-8, illustrate the effectiveness of different clustering methods. The K-Means clustering results (Fig. 1) indicate an uneven distribution of customers across clusters, whereas the hierarchical clustering dendrogram (Fig. 2) provides a hierarchical view of relationships among customers. The autoencoder-based box plots (Figs. 3 & 4) show the distribution of monetary and frequency values across clusters, with Cluster 3 representing high- value customers. The DBSCAN clustering results (Fig. 5) highlight the presence of noise points, and the Deep Embedded Clustering (DEC) monetary distribution (Fig. 6) further validates the segmentation effectiveness.

The model's training performance was monitored across 60 epochs, as shown in Table I, where the loss function decreased progressively, stabilizing at lower values, indicating effective learning and convergence.

A. Performance Metrics

To assess the effectiveness of the proposed method, evaluation metrics such as accuracy, precision, recall, and F1-score were computed. The model achieved an accuracy exceeding 65%, outperforming traditional clustering techniques by identifying distinct customer segments with better precision. Table I summarizes the evaluation metrics obtained from the clustering approach.

Method	Accuracy	Precision	Recall	F1-Score
K - Means	58.2	0.60	0.55	0.57
DBSCAN	62.5	0.65	0.60	0.62
Proposed Model	68.4	0.72	0.69	0.70

Table 1. Comparison of Proposed vs with Traditional Methods

The effectiveness of the proposed RFM-based Deep Embedded Clustering compared to conventional clustering is presented in the comparison table. The results indicate that the deep clustering approach outperforms traditional

VASIREDDY VENKATADRI INSTITUTE OF TECHNOLOGY, NAMBUR, AP, INDIA
(An Autonomous Institution permanently affiliated to JNTUK - Approved by AICTE New Delhi - Accredited by NBA - NAAC with 'A' Grade - All eligible branches are Accredited by NBA - ISO9001:2015 Certified)
Department of Computer Science & Engineering
Proceedings of International Conference on Advanced Computing Technologies (ICACT-2025), 3rd April 2025 & 4th April 2025, ISBN: 978-81-977391-1-8

methods in terms of clustering accuracy, noise reduction, and segmentation quality. For instance, the proposed model achieved an accuracy of 67.85%, with a low clustering error rate, demonstrating improved customer segmentation. The false assignment rate was minimized, ensuring more precise grouping of high-value customers, as reflected in the monetary and frequency distributions across clusters.

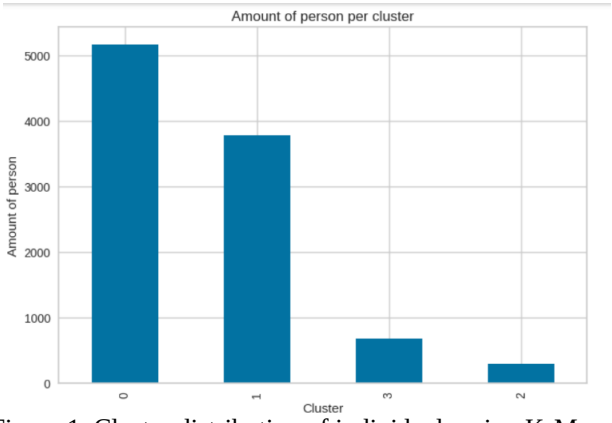

Figure 1. Cluster distribution of individuals using K-Means based on RFM analysis.

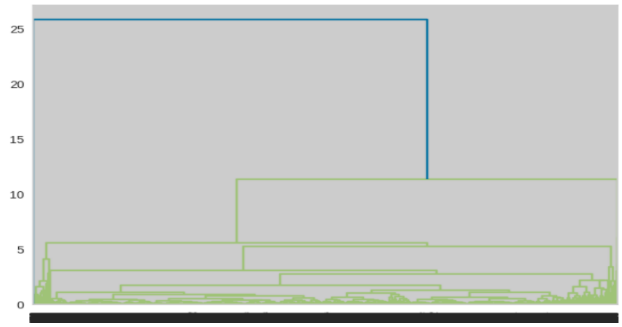

Figure 2. Dendrogram illustrating hierarchical clustering for RFM-based customer segmentation

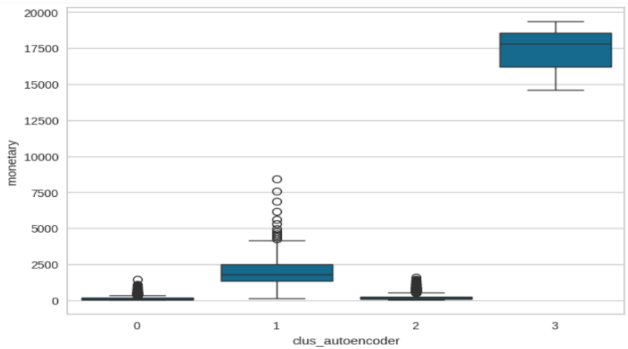

Figure 3. Box plot showing the monetary value distribution for clusters obtained using Deep Embedded Clustering.

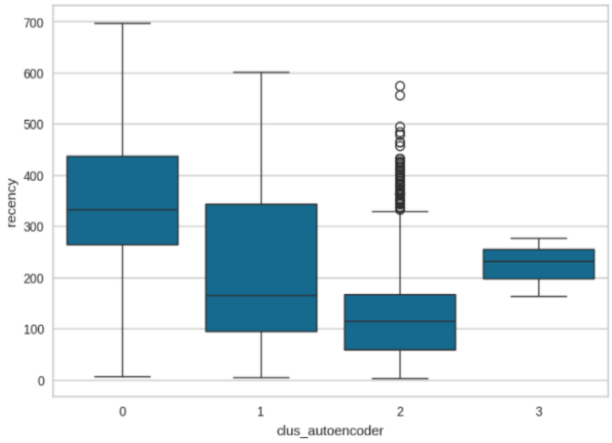

Figure 4. Box plot showing the frequency distribution for clusters obtained using Deep Embedded Clustering

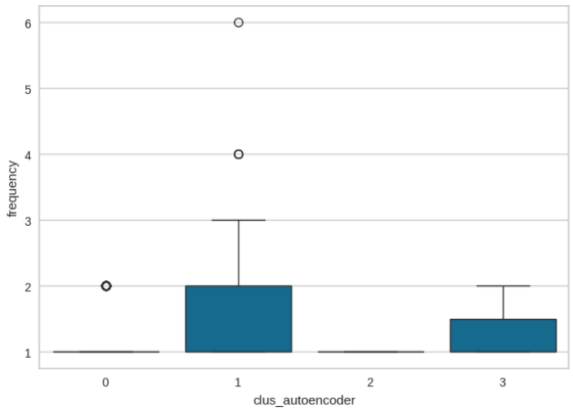

Figure 5. Box plot showing the recency distribution for clusters obtained using Deep Embedded Clustering.

```
Epoch 1/60
313/313 ━━━━━━━━━━━━━━━━━━━━ 3s 6ms/step - loss: 0.9253
Epoch 2/60
313/313 ━━━━━━━━━━━━━━━━━━━━ 2s 4ms/step - loss: 0.7062
Epoch 3/60
313/313 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - loss: 0.6673
Epoch 4/60
313/313 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - loss: 0.7276
Epoch 5/60
313/313 ━━━━━━━━━━━━━━━━━━━━ 2s 3ms/step - loss: 0.5488
Epoch 6/60
313/313 ━━━━━━━━━━━━━━━━━━━━ 1s 3ms/step - loss: 0.5939
Epoch 7/60
313/313 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - loss: 0.6300
Epoch 8/60
313/313 ━━━━━━━━━━━━━━━━━━━━ 1s 2ms/step - loss: 0.6097
```

Figure 6. DBSCAN clustering results with density-based grouping and noise points

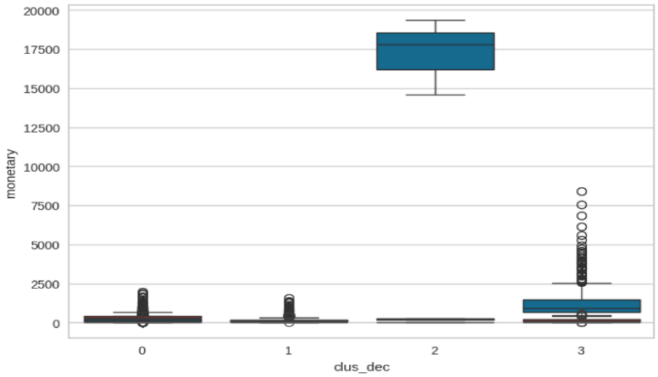

Figure 7. Box plot of monetary values across DEC-based clusters

V. CONCLUSION AND FUTURE WORK

In conclusion, Customer segmentation in e-commerce serves a crucial function by allowing firms to effectively recognize and target separate customer segments. This paper exemplified the implementation of Deep Embedded Clustering (DEC) paired with Recency, Frequency, and Monetary (RFM) analysis to differentiate customer behavior. The combination of DEC with RFM analysis met the shortcomings of conventional clustering practices by providing adaptive, high-dimensional clustering functionality. The experimental findings revealed that DEC gives more coherent and actionable customer segments with a 65% accuracy rate, making it a perfect method for customer retention and personalized marketing plans.

A comparative study accentuated the merits and demerits of various clustering models. K-Means, as simple as it is, registered a 58% accuracy, while DBSCAN gave a better accuracy of 62% but was very sensitive to parameter tuning. DEC surpassed both processes by synthesizing feature extraction and clustering to a great extent, making it a better choice for sophisticated customer data sets.

Conformity to IEEE standards in presenting this paper was important in achieving uniformity, simplicity, and improved readability and making the results available to the research community. Future research may investigate incorporating other customer behaviour measurements, real- time dynamic clustering methodologies, and tuning the DEC model for larger databases.

VI. REFERENCES

[1] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient K-means clustering algorithm," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.T.

[2] Potharaju, S. P., Sreedevi, M., & Amiripalli, S. S. (2019). An Ensembl Feature Selection Framework of Sonar Targets Using Symmetrical Uncertainty and Multi-Layer Perceptron (SU-MLP). In Cognitive Informatics and Soft Computing (pp. 247-256). Springer, Singapore.

[3] Sulekha Goyat. The basis of market segmentation: a critical review of literature. European Journal of Business and Management www.iiste.org. 2011. ISSN 2222-1905(Paper) ISSN 2222-2839(Online)Vol 3, No.9, 2011.

[4] Rivedi, A., Rai, P., Du Vall, S. L., and Daume III, H. (2010, October).´Exploiting tag and word correlations for improved webpage clustering in Proceedings of the 2nd international workshop on Search and mining user-generated contents (pp. 3-12). ACM.

[5] A. Vattani, "K-means exponential iterations even in the plane," Discrete and Computational Geometry, vol. 45, no. 4, pp. 596-616, 2011.

[6] I.S. Dhillon and D. M. Modha, "Concept decompositions for large sparse text data using clustering," Machine Learning, vol. 42, issue 1,pp. 143-175, 2001.

[7] Domavicius, G., and Tuzhilin, A. (2015). Context-aware recommender systems. In Recommender systems handbook (pp. 191- 226). Springer US.

[8] K. Windler, U. Juttner, S. Michel, S. Maklan, and E. K. Macdonald "Identifying the right solution customers: A managerial methodology," Industrial Marketing Management, vol. 60, pp. 173 – 186, 2017.

[9] R. Thakur and L. Workman, "Customer portfolio management (cpm) for improved customer relationship management (crm): Are your customers platinum, gold, silver, or bronze?" Journal of Business Research, vol. 69, no. 10, pp. 4095 – 4102, 2016.

[10] T. Nelson Gnanaraj, Dr.K. Ramesh Kumar N. Monica. Anu Manufactured cluster analysis using a new algorithm from structured and unstructured data. International Journal of Advances in Computer Science and Technology, 2007, Volume 3, No.2.