Privacy Preserving Federated Graph Learning with Data Envelopment Analysis Driven Interpretable Customer Segmentation Framework

Pentareddy Ashalatha[1], Dr. G. Krishna Mohan[2]
Research Scholar, Department of Computer Science & Applications, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India[1]
Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India[2]

Abstract—**Customer segmentation plays a critical role in retail analytics by enabling personalized marketing, optimized resource allocation, and data-driven strategic decision-making. However, customer data is often distributed across multiple retail branches and contains sensitive transactional information, creating significant challenges related to privacy preservation, regulatory compliance, and model interpretability. Traditional segmentation approaches, including clustering algorithms and centralized deep learning methods, typically require aggregated data storage, which increases privacy risks and limits secure deployment in distributed retail environments. In addition, many existing methods fail to capture complex relational patterns such as co-purchasing behavior and inter-customer structural dependencies. To address these limitations, this study proposes FedGraph-DEA, a novel hybrid framework integrating federated learning, graph neural networks (GNNs), and Data Envelopment Analysis (DEA) for privacy-preserving and efficiency-aware customer segmentation. The framework first employs Distributed Federated Convolutional Autoencoders to extract latent customer representations from decentralized retail datasets. Similarity graphs are then constructed locally, followed by federated GNN-based community detection to identify structurally coherent customer groups without sharing raw data. Finally, DEA is applied to evaluate the operational efficiency of the discovered customer segments. Experimental evaluation was conducted using the UCI Online Retail dataset partitioned across five simulated non-IID client nodes. The proposed model achieved 96.1% accuracy, 0.95 precision, 0.96 recall, and 0.95 F1-score when compared with pseudo-ground-truth clusters generated through K-Means reference clustering. Furthermore, the framework obtained a silhouette score of 0.74 and a modularity value of 0.57, demonstrating strong cluster compactness and structural separation. The proposed system provides a scalable, interpretable, and privacy-preserving solution for distributed retail analytics.**

Keywords—Customer segmentation; federated learning; graph neural networks; data envelopment analysis; privacy preservation

I. Introduction

In today's digital economy, customer segmentation is generally recognized as a pillar of data-driven decision-making and marketing [1], [2]. By segmenting customers into substantive groups, companies can tailor services to individuals, improve resource usage, and create targeted campaigns [3]. Nevertheless, with the exponential growth of online transactions and widespread sensitive customer data proliferation [4], privacy-enhancing and explainable methods for segmentation have become ever more important [5]. Regulatory protocols like GDPR and HIPAA mandate rigorous adherence to privacy principles, as businesses at the same time seek transparent and actionable information from their analytics platforms [6]. Conventional centralized approaches leave organizations vulnerable to data leakage, breaches, and regulatory non-compliance, and thus, there is a need for paradigms of federated learning [7]. Of equal significance is interpretability, as stakeholders and managers must not only know what segments exist but also the reason why those segments are effective and worthwhile [8]. In this context, privacy-preserving interpretable customer segmentation has become a salient area of attention in academia and the industry alike, filling the gap between secure distributed learning and decision-driven insights [9].

Previous approaches to customer segmentation have primarily utilized statistical clustering algorithms like K-Means, DBSCAN [10], [11], and hierarchical clustering, or machine learning classifiers like Random Forests and XGBoost [12], [13]. Although these approaches provide computational ease, the models tend to lack flexibility in accommodating high-dimensional transactional data and are not robust in distributed settings [14]. In addition, the model tends to need centralized data collation, which presents privacy threats in areas like finance, healthcare, and e-commerce [15]. Autoencoders based on deep learning have also recently been considered for extracting latent behavior features; however, without federated approaches, the models still need to pool raw data at a central server[16]. In terms of interpretability, clustering and classification methods usually return black-box outputs, providing decision-makers with little understanding of the efficiency or strategic importance of the segmented findings [17]. Therefore, current frameworks fail to meet the two-pronged challenge: protecting sensitive customer information in decentralized ecosystems and returning interpretable, transparent results for managers [18], [19].

The FedGraph-DEA model plan creates a customer segmentation framework that is privacy-protected and interpretable, and is designed to work in a distributed retail-setting. DFCA is used to learn high-quality latent behavioral representations using transactional data that is decentralized without the exchange of raw information. These similarity embeddings are organized into customer similarity graphs in order to obtain co-purchase dependencies and multi-hop relational patterns. With the help of Federated Graph Neural Networks (GNNs), the recognition of coherent communities of customers in non-IID data distributions is then realized. Lastly, Data Envelopment Analysis (DEA) measures the operational efficiency of every segment, which allows for actionable performance-based and interpretable segmentation results by managers.

A. Research Motivation

The rise of the distributed retail systems has escalated the necessity of customer segmentation strategies that adequately protect data privacy and yet express intricate behavioral relationships [20]. Conventional centralized systems reveal sensitive customer data and do not represent relational relationships like co-purchase behavior. Simultaneously, the existing models are usually not interpretable, which constrains their applicability in managerial decision-making. The rationale behind this research is to come up with a single framework, which is able to take care of privacy preservation, relational intelligence, and transparency, so as to institute the correct and practical customer segmentation in the retail settings where data is sensitive and decentralized.

B. Research Significance

This research is significant because it attempts to fill vital shortcomings of current customer segmentation methods by combining the concepts of privacy protection, relational modelling, and interpretability under a single system. Contrary to the old-fashioned approaches, it provides a secure data processing in distributed contexts, as well as, the transformation of intricate customer interactions and offers efficiency-related insights to make a decision. The research has also helped the academic research as well as practical applications through facilitating scalable, transparent, and performance-driven analytics. It can be applied to privacy sensitive areas, like store, finance and healthcare where data analysis requires the use of secure and readable data.

C. Problem Statement

There are several limitations associated with the current techniques of customer segmentation in a distributed retail setting. Firstly, current traditional approaches of customer segmentation tend to aggregate private and sensitive information about customers, exposing the system to higher risks of data leaks and non-conformity with relevant regulations [21]. Finally, the majority of current models used for customer segmentation have no interpretability; consequently, they provide little information about the factors that affect customer segmentation [22]. In spite of the attempts by modern researchers to introduce federated learning models to address privacy and relational intelligence challenges [23], interpretability and efficiency-based assessment of customer segments remain underestimated aspects. Thus, the development of a unified approach to customer segmentation seems to be essential.

D. Key Contribution

Introduced a customer segmentation algorithm that maintains customer privacy to solve the problem of data confidentiality, scalability, and distributed retail environment constraints as well as performs behavioral representation learning without any data aggregation in the centralized setting.

Provides the use of DEA as an additional post-segmentation evaluation tool to measure the efficiency of the graph-based customer communities formed based on operational efficiency.

Proposes a privacy-preserving federated similarity graph formation mechanism to capture multi-hop dependency among the customers' behavior patterns in non-IID distributed learning environment without sharing raw customer data among the retailers.

Combines Federated Graph Neural Network methods with Distributed Federated Convolutional Autoencoder (DFCA), allowing joint modeling of compressed behaviors and relations in the decentralized retail setting demonstrates that the combined DFCA + Federated GNN + DEA framework improves segmentation quality, clustering coherence, and interpretability compared to conventional centralized and federated baseline models.

The research is presented in the following structure: Existing studies are presented in Section II. Section III covers problems with the existing study. The suggested methodology is presented in Section IV. Section V includes the results and a discussion of this research. Conclusions and future works are covered in Section VI.

II. Related Work

Le et al. in [24] presents PersonalFR, a personalized federated recommender system that utilizes autoencoder-based models to learn user-item interactions without revealing raw data. The local updating of encoders and aggregation of decoders on the server side maintains data privacy. The experiments on real-world datasets show that PersonalFR offers performance close to centralized models and with lower communication overhead. The non-interpretability of the autoencoder's latent features also restricts the explainability of the segmentation outcomes. J. M. John, O. Shobayo, and B. Ogunleye in [25] compares different clustering algorithms—K-means, Gaussian Mixture Models (GMM), DBSCAN, Agglomerative Clustering, and BIRCH—on a UK-based online retail dataset with more than 500,000 records. Based on the RFM (Recency, Frequency, and Monetary) framework, the study seeks to improve decision-making in the retail industry. Nonetheless, the use of centralized data processing in the study creates data privacy and scalability issues. M. Alves Gomes and T. Meisen in [26] evaluates the 105 publications published between 2000 and 2022. The four stages of the process that are identified in the work are data collection, customer representation, segmentation, and targeting. The integration of sophisticated approaches such as PCA and SOMs is also described in the review.

Wang in [27] investigates the coupling of deep learning models, customer segmentation, online marketing, and swarm intelligence algorithm models. The synergy aims at maximizing the performance of the clustering by the global search capability of swarm intelligence. The study does not address the problem of data heterogeneity in distributed environments. In addition, the interpretability of the resulting customer segments is constrained, as well as integration with efficiency assessment frameworks.

A. Sharma, N. Patel, and R. Gupta in [28] investigates the convergence of deep learning methods with predictive analytics to enhance customer segmentation techniques. Through the use of neural networks to examine intricate customer data, the study seeks to reveal subtle behavioral patterns that may be missed by conventional methods. Nevertheless, the study admits difficulties in model interpretability and the necessity for large computational resources, which could restrict its usage for smaller businesses or companies with limited technical infrastructure. Y. Long et al.,[29] Presented a federated deep clustering method for retail customer segmentation, integrating local autoencoders with a central K-means clustering model. Using federated learning, their model ensured customer data privacy across different branches while maintaining an overall segmentation accuracy of around 88%. But the study was not concerned with the evaluation of the resulting clusters in terms of operational efficiency or business value, which curtails its applicability in strategic decision-making. Contrarily, the current FCA–DEA framework not only improves privacy via federated training but also encompasses a DEA-based layer to evaluate and rank customer segments, which fills the gap between high-performance clustering and business interpretability in real-world scenarios.

P. Hu et al., [30] proposed a federated CNNs-based privacy-aware clustering model to partition customer data scattered over multiple retail stores. Their approach effectively circumvented raw data centralization and had a clustering accuracy of approximately 90%, which proved the feasibility of federated deep learning for customer analytics. However, the approach did not incorporate an interpretability function to justify the relevance of the clusters and provided no facility for benchmarking the operating efficiency of every segment. This limitation justifies the necessity for hybrid models such as the FCA–DEA framework, which pairs federated unsupervised learning with a post-segmentation DEA module for clear and actionable efficiency assessment.

Y. Jiang et al., [31] introduced federated VAE to use in distributed customers profiling in the retail industry. At their current stage their approach employed VAEs in order to elicit non-linear latent representations of customer purchasing tendency subject to privacy-respecting constraints with a reconstruction precision of approximately 92.1%. Although the model has shown robust features learning without focusing sensitive data, still, it did not consider the problem of making sense of the latent features to apply business results, as well as assessing the relative efficiency of the generated groups of customers. In comparison, the suggested FCA-DEA framework integrates deep representation learning with operational efficiency ranking of DEA so that businesses could translate latent features into plain, usable forms in terms of resource optimization and targeted marketing.

Based on the reviewed literature, it can be seen that past research either concentrated on centralized clustering and deep learning processes, which ignore privacy, or federated models, which do not analyze interpretability or operational efficiency. The graph-based model, though effective in depicting feature interaction still uses centralized access to data. The FedGraph-DEA framework will overcome such shortcomings by combining federated autoencoders to learn decentralized representations, graph neural networks to detect communities, and data envelopment analysis to identify the efficiency of the model. This is a hybrid design that provides secure learning, increased segmentation accuracy, and transparent decision support, something that previous models did not have.

III. FedGraph-DEA Framework for Customer Segmentation

The FedGraph-DEA method employs the concept of a single pipeline that combines decentralization representation learning, graph-based feature interaction model, and efficiency-based interpretability. Every client node separately trains a Distributed Federated Convolutional Autoencoder (DFCA) to learn the customer latent embeddings without having access to raw data. These embeddings are then used to create a customer similarity graph based on a cosine affinity. The graph establishes the connectivity of the nodes and is enhanced using federated Graph Neural Network (GNN), which reflects on multi hop customer relations through federated weight aggregation across distributed subgraphs. Communities that are formed as a result are customer segments. These segments are then measured by Data Envelopment Analysis (DEA) wherein each of the segments is considered as a decision-making unit and the evaluation is done in terms of business efficiency by considering the marketing inputs and profit-related outputs. DFCA induced graph formation, GNN-based segmentation and DEA-based evaluation are new to interact. FedGraph-DEA is the only method where privacy preservation, relational intelligence, and interpretable efficiency assessment are integrated into a single framework in contrast to the current federated clustering, federated GNN, or single method of standalone DEA. Fig. 1 shows the overall workflow.

The innovation introduced by the FedGraph-DEA model is its comprehensive integration of three distinct yet complementary approaches in one privacy-preserving segmentation architecture for distributed retailers. In contrast to the traditional federated clustering models that concentrate exclusively on privacy-preserving capabilities, this new model is capable of analyzing higher-order customer relations with the aid of graph neural networks while concurrently measuring the efficiency of their operations with DEA.

The principal novelty of this methodology lies in the unified integration of federated representation learning, graph-based relational modeling, and efficiency-driven evaluation within a single decentralized analytical framework. Unlike conventional customer segmentation approaches that rely on centralized data aggregation and purely distance-based clustering, this architecture enables privacy-preserving latent feature extraction through federated convolutional autoencoders while simultaneously capturing higher-order behavioral dependencies via graph neural networks. The transformation of learned embeddings into customer similarity graphs allows structural and multi-hop relationships to be incorporated into segmentation, addressing limitations of attribute-only models. Furthermore, the incorporation of Data Envelopment Analysis introduces an operational efficiency perspective that links discovered segments to measurable input–output performance metrics, enhancing managerial interpretability. This synergistic combination of privacy preservation, relational intelligence, and efficiency-based interpretability constitutes a novel contribution to distributed customer analytics research.

Fig. 1 illustrates the FedGraph-DEA approach for customer segmentation under the privacy-preserving setting. First, the UCI Online Retail dataset is considered, then the preprocessing phase involves data cleaning, normalization, and feature engineering. The next layers, known as federated learning, employ distributed autoencoder methods and secure aggregation to learn the latent space representation without sharing data. Using the embeddings, a graph structure can be built, which is utilized by federated graph neural networks for community detection. Finally, DEA is employed to analyze efficiency.

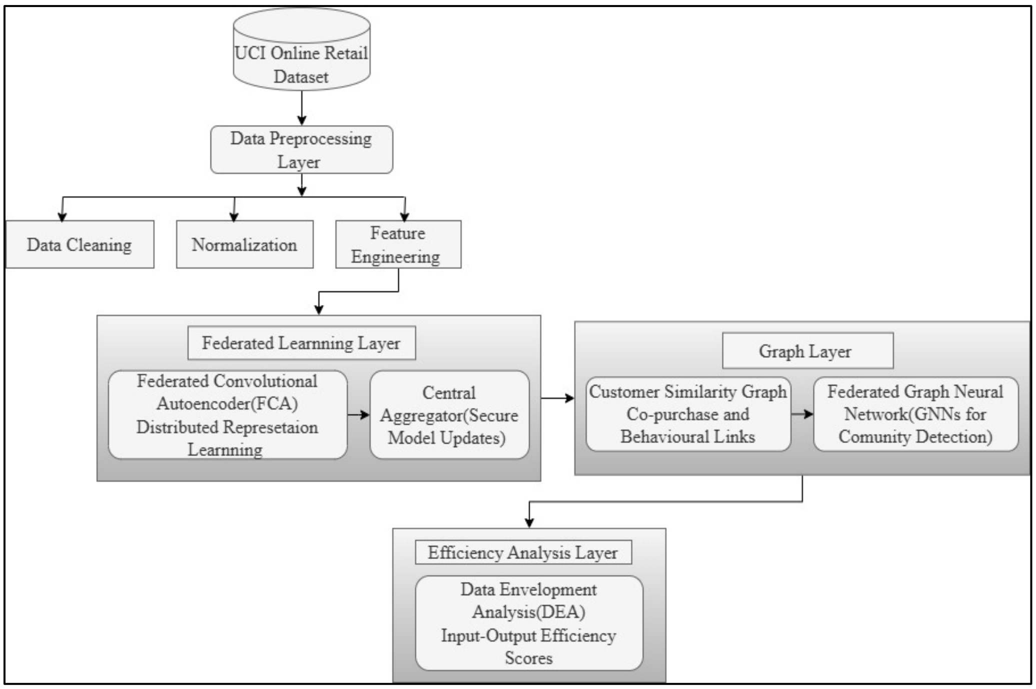

Fig. 1. Privacy-preserving federated graph-based customer segmentation workflow.

A. Data Collection

UCI Online Retail is an online retail dataset that entailed in excess of 500,000 transaction records in various nations in the 2010-2011 [32]. It stores specific buying behavior such as invoices, product characteristics, amount, costs, time values, customer names, and addresses. It is appropriate in terms of customer segmentation, federated learning, privacy-preserving analytics, and RFM behavioral modeling for distributed retail operations due to its heterogeneity and realistic nature. In the case of federated simulation, the database was split among five clients according to the country attribute for transactions, thus leading to naturally non-IID distributions, as buying habits and preferences vary greatly depending on location. This division allows the creation of distinct local databases for each client in an actual federated manner without the need for more than one data source physically.

Table I illustrates five sample transactions showing product codes, descriptions, quantities, and prices, demonstrating the dataset’s structure and how individual purchase records form the basis for feature engineering and segmentation.

TABLE I. SAMPLE TRANSACTION RECORDS FROM THE UCI ONLINE RETAIL DATASET

Invoice No	Stock Code	Description	Quantity	Unit Price
536365	85123A	White Mug	6	2.55
536366	71053	Floral Notebook	3	4.95
536367	84406B	Pink Hanging Bag	2	6.95
536368	84029G	Heart Decoration	8	1.25
536369	22960	Jam Making Set	1	11.95

B. Data Preprocessing

Customer segmentation requires strict data preprocessing as it converts raw transactions into processed and meaningful patterns of customer behavior. This is essential in eliminating noise, ensuring data quality, and facilitating proper, reliable segmentation analysis.

1) Data cleaning: Missing Customer IDs will be removed and negative values will be deleted because that means returns or cancellations rather than purchases. Data integrity is ensured by discarding duplicate invoices. The resulting set of data only contains valid and applicable transactions.

2) Feature engineering: The highly set Recency-Frequency-Monetary (RFM) is used to extract customer-level behavioral attributes. The Eq. (1) and Eq. (2) define the calculations performed for each customer i,

$$R_i = t_{ref} - t_{last\ purch,i} \tag{1}$$

$$M_i = \sum_{j=1}^{n_i} p_{i,j} \tag{2}$$

where, R_i represents the period of time since the time of the last purchase made by the customer, F_i is the total number of purchases made, and M_i is the cumulative purchase value.

3) Normalization: Min-max normalization is used to rescales all numerical features, which ensure that the magnitude of the numerical feature is similar and scale is not dominant in federated and graph-based learning computed using Eq. (3):

$$X_{scaled} = \frac{.X - X_{min}}{X_{max} - X_{min}} \tag{3}$$

where, X_scaled is the normalized feature value, X is the original data point, while X_{min} and X_{max} denote the minimum and maximum feature values used for scaling.

C. Distributed Federated Convolutional Autoencoder (DFCA)

The DFCA is the first module in the FedGraph-DEA framework and is used for training privacy-preserving latent embeddings for customers' transactional behaviors. Given that customer transactions are distributed over different branches and that their centralization is prohibited by privacy laws, each client will train an autoencoder using their normalized RFM feature vectors. Here, the encoder maps the feature vector into the embedded representation, while the decoder tries to reconstruct the feature vector from this latent embedding. The encoder working is determined as Eq. (4):

$$h_i^{(k)} = \sigma(W_{e,k} * x_i + b_{e,k}) \tag{4}$$

where $x_i \in \mathbb{R}^d$denotes the normalized RFM feature vector of customer i, $W_{e,k}$and $b_{e,k}$represent the encoder weight matrix and bias vector of client k, respectively, and $\sigma(\cdot)$ denotes a nonlinear activation function such as ReLU or sigmoid. The encoder is able to compress high-dimensional information about behavior into low-dimensional representation that is capable of preserving purchasing attributes despite noise and redundancy. The latent representation is subsequently reconstructed through the decoder network to preserve reconstruction fidelity, as expressed in Eq. (5):

$$\hat{x}_i^{(k)} = \sigma(W_{d,k} * h_i^{(k)} + b_{d,k}) \tag{5}$$

where $\hat{x}_i^{(k)}$ denotes the reconstructed customer feature vector, while $W_{d,k}$and $b_{d,k}$correspond to decoder parameters for client k. The decoder attempts to minimize information loss between the original and reconstructed customer behavior representations. The DFCA model is optimized using reconstruction loss, which measures the difference between the original and reconstructed feature vectors is given in Eq. (6),

$$\mathcal{L}_{\mathrm{rc}(k)}^{k} = \frac{1}{n_k} \sum_{i=1}^{n_k} |x_i - \hat{x}_i^{(k)}|_2^2 \tag{6}$$

where, N_k represents the number of customer samples available at client k. This objective ensures accurate preservation of local behavioral patterns while learning generalized latent embeddings suitable for federated graph construction and downstream segmentation tasks. In order to encourage the stability of models with highly non-IID purchasing distributions, and L_2 regularization factor can be defined as Eq. (7):

$$\mathcal{L}_{\mathrm{AE}}^{(k)} = (\mathcal{L}_{\mathrm{rc}}^{k} + \lambda|\theta_k|_2^2) \tag{7}$$

where, θ_k and ϕ_k represent decoder and encoder parameters, and λ controls the fine of huge weights. The $\theta_k = \{W_{e,k}, b_{e,k}, W_{d,k}, b_{d,k}\}$ refer to all the trainable parameters. The learning objective worked out in Eq. (8):

$$\min_{\theta_k} = \mathcal{L}_{AE}^{(k)} \tag{8}$$

where, θ denot $\mathcal{L}_{AE}^{(k)}$ represents its reconstruction loss; the objective minimizes this loss to learn optimal local latent representations from client data. This will guarantee uniform extraction of latent purchasing representations between different branches, stores or regions.

After local training, model parameters (but not raw data) are sent to an international aggregation server that is secure and is located centrally. The Federated Averaging (FedAvg) algorithm determines which client models are used in this process by means of a weighted average of the model whose dataset is the largest to the smallest. The international brief at the communication round $t + 1$ is expressed can be defined as Eq. (9):

$$w^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{n_{\mathrm{total}}} w_k^{(t)} \tag{9}$$

In this averaging process, $w_k^{(t)}$ are the local model weights of client k at the t th iteration. n_k is the size of the data sample of each node, and n total is the sum of all the samples. When aggregation has been done the server transmits the new global model to all the clients as received in Eq. (10):

$$\theta_k^{(t+1)} \leftarrow w^{(t+1)}, \ n_{\mathrm{total}} = \sum_{k=1}^{K} n_k \tag{10}$$

where, $\theta_k^{(t+1)}$ denotes updated local parameters for client k, $w^{(t+1)}$ is the aggregated global model, n_k is client data size, and nn_{total}represents the total samples.

The synchronized update scheme allows converging to a common embedding space even in cases of statistical heterogeneity of distributed datasets. FedAvg is preferred because it is efficient in communication, does not suffer from instabilities in non-IID environments, and is scalable in industrial federated systems. Notably, the exchange of model updates only is provided, and the raw data do not leave their locations, which reduces the privacy leakage and ensures that the enterprise and regulatory data-protection policies are followed. The framework provides a uniform embedding space to clients through a combination of local representation learning with DFCA and global alignment with FedAvg. The resulting embeddings are privacy-preserving, statistically resistant and behaviorally meaningful, which makes them a solid basis of building customer similarity graphs and makes them useful in graph-based community detection in the FedGraph-DEA architecture.

D. Customer Similarity Graph Construction

Once latent vectors are created by DFCA, the step that follows is to create a customer similarity graph that reflects behavioral relations among customers. Customers are represented as node and the relationships are created between customers with a high similarity in the latent embeddings. Similarity of behavior between embeddings is done through cosine similarity. A threshold parameter is used to filter weak or unimportant connections and only on useful relationships. The latent embedding space is converted into a structured graph of co-purchase behavior and dependency by this process. The resulting graph forms the basis of the Federated Graph Neural Network, which allows the learning of higher-order interaction with customers, without compromising the decentralized and privacy-preserving federated architecture, as stated in Eq. (11):

$$S_{ij} = \frac{z_i \cdot z_j}{|z_i|_2 |z_j|_2} \tag{11}$$

In this cosine similarity function, z_i and z_j are latent feature embeddings for customers i and j, and $|_2$ denotes the L2 norm. S_{ij} measures the degree of behavioral similarity between two customers in embedding space. This module constructs a customer similarity graph, where nodes represent customers and edges indicate behavioral similarity and the similarity of their behavior being edges. This research reveals the co-purchase patterns and latent association among distributed stores and creates a structured relational network which feeds the federated GNN to detect communities and segment the graph with respect to it.

E. Federated Graph Neural Networks (GNNs)

The Federated Graph Neural Network (GNN) is utilized on the customer similarity graphs of all clients to discover the community structures as well as latent relational patterns. A local GNN is trained on each subgraph of the client, with a node taking information on its connected neighbors. By layer wise propagation, node representations undergo an update process based on averaged or weighted features of their near feature, bringing in greater order dependence on feature interactions. Once the local training is made, just GNN weight matrices are exchanged with a central server to aggregate without exposing raw data or making them public. The global GNN sum total is used to bring the relational representations of the clients without direct data exchange. This procedure makes it possible to detect coherent communities without compromising privacy, as expressed can be defined as Eq. (12).

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ij}} W^{(l)} h_j^{(l)}\right) \tag{12}$$

Here, $h_i^{(l+1)}$ denotes the node representation of customer i at GNN layer l, and $\mathcal{N}(i)$ is its neighborhood set. The normalization factor c_{ij} scales neighboring influences, while $W^{(l)}$ and activation σ update embeddings derived in Eq. (13),:

$$W_{\text{global}}^{(l)} = \sum_{k=1}^{K} \frac{n_k}{n_{\text{total}}} W_k^{(l)} \tag{13}$$

Eq. (10) defines federated weight aggregation for the GNN. $W_k^{(l)}$ are layer-wise parameters trained locally on each client graph, and $W_{\text{global}}^{(l)}$ is the aggregated matrix ensuring coherent feature propagation across distributed graphs. The Fed-GNN generates community-aware embeddings by aggregating neighborhood information from each individual client across distributed subgraphs without requiring data centralization. These embeddings reveal buying communities and latent behavioral addiction, and create strong, graph-structured clusters, upon which interpretable and relationally grounded buyer segmentation is built.

F. Data Envelopment Analysis (DEA) and Decision Support

Once the Fed-GNN determines customer segments or communities, DEA determines the relative efficiency of each of the segments as a decision-making unit (DMU). DEA is useful in comparing several inputs, including marketing cost, campaign frequency and acquisition expense against outputs, including profit, retention and purchase value. It finds an efficiency score for each segment between 0 and 1 by finding a solution to a linear programming model. These segments with a score of 1 are considered efficient which means that the model has the best resource to return conversion and those that score below are given improvement targets. DEA will convert the abstract GNN clusters into real world business knowledge and connect data-driven segmentation with the operational efficiency. It is this interpretive layer that makes FedGraph-DEA not just predictive in nature. It was mentioned in Eq. (14):

$$\text{s.t} \; \frac{\sum_{r=1}^{R} u_r y_{jr}}{\sum_{s=1}^{S} v_s x_{js}} \leq 1, u_r, v_s \geq 0 \tag{14}$$

This constraint ensures that all decision-making units j operate within an efficiency frontier ≤ 1. The non-negativity of u_r, v_s maintains the economic interpretability of weights, enforcing DEA feasibility. DEA produces a score on the efficiency of every customer group, which measures how profitable the proposed framework is in comparison to the marketing resources. This is the proposed research result in ranking of the effective and ineffective segments and gives the clear managerial information on the budget redistribution, focused campaigns, and performance recovery across the distributed branches.

G. Integrated Optimization Objective

The FedGraph-DEA also includes the last stage as a way of integrating all the previous functionalities into one overall optimization goal. The overall loss can be defined as the sum of three components: 1) DFCA reconstruction loss which guarantees that the model can represent the behavioral aspects accurately, 2) the Fed-GNN relational loss which generates inter-customer dependencies and 3) the DEA-based efficiency regularization term, which encourages interpretability and business relevance. A tradeoff between these goals enables the model to learn jointly, correct, relational and interpretable segmentations. The coefficients α, β, γ The three dimensions, α, β, γ, and three are used to control the trade-off between reconstruction, preservation of community structure and maximization of efficiency. Optimizing this composite loss in this study provides customer segments which are not only statistically valid but also economically relevant thus forming a holistic learning-and-decision model of distributed segmentation. The joint optimization is the result of a balanced global model that considers accuracy, relational integrity and business interpretability. Reducing this loss study produces the end FedGraph-DEA system that provides privacy-preserving, community-conscious, and efficiency interpretable customer segmentation that can be deployed at the enterprise level.

A distributed federated learning architecture that is used to achieve privacy-protecting and interpretable segmentation of customers. It starts with customer data in the form of transactional and RFM, which is handled using a Distributed Federated Convolutional Autoencoder (DFCA) to captures the compressed latent behavioral representations without compromising data confidentiality. The encoder is used to capture meaningful hidden patterns and the decoder is used to guarantee reconstruction fidelity with federated aggregation permitting a secure weight sharing rather than a raw data sharing. The learned embeddings are then converted into a customer similarity graph, where the nodes are the customers and the edges are the behavioral relations between customers based on the co-purchase and similarity scores. The Federated Graph Neural Networks (GNNs) are subsequently utilized to do local subgraph learning and global parameter aggregation, which helps to perform strong community discovery in decentralized and non-IID conditions. The end result will be structurally coherent, efficiency-focused, and interpretable by the manager, customer segments that can be distributed in retail analytics. The model architecture diagram as shown in Fig. 2.

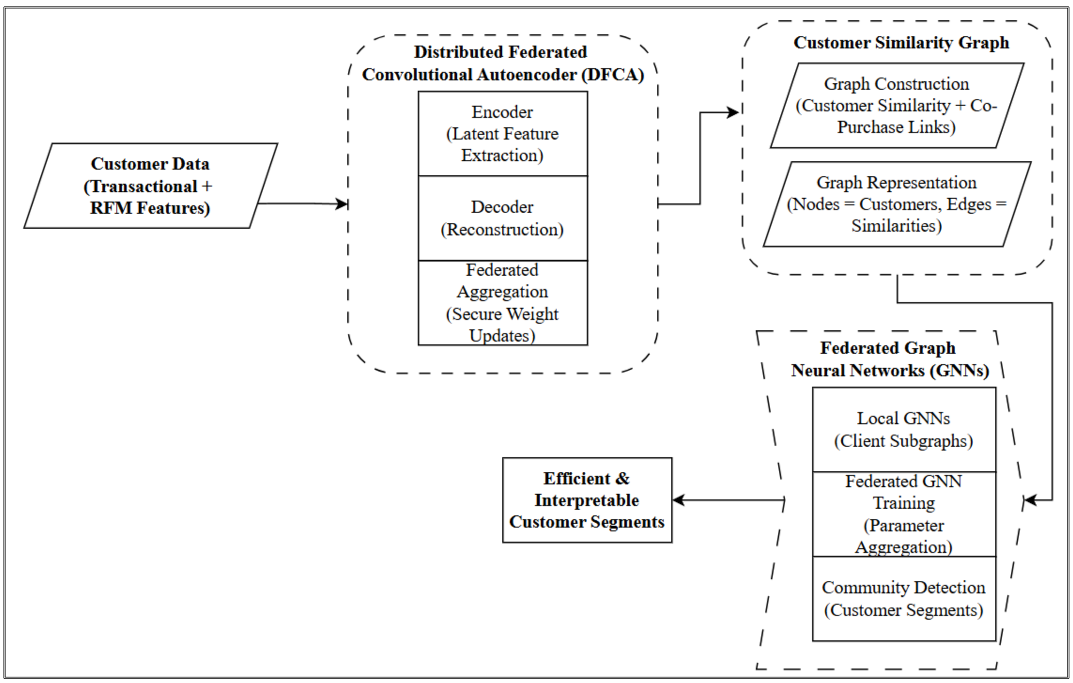

Fig. 2. Federated graph learning pipeline for customer segmentation.

Fig. 2 illustrates a federated learning-based pipeline for customer segmentation that integrates representation learning and relational modeling. Customer transactional and RFM data are processed using a Distributed Federated Convolutional Autoencoder (DFCA), where local encoders extract latent features and decoders ensure reconstruction fidelity, while model parameters are securely aggregated across clients. The learned embeddings are used to construct a customer similarity graph capturing behavioral relationships. Federated Graph Neural Networks (GNNs) then perform decentralized training on local subgraphs, enabling community detection. This process produces privacy-preserving, structurally meaningful, and interpretable customer segments for effective decision-making.

Algorithm 1 securely trains customer patterns on distributed nodes, constructs a similarity graph, identifies relational communities using federated GNNs, and optimizes business efficiency of each segment via DEA to provide the interpretable and privacy-enforceable segmentation.

Algorithm 1: FedGraph-DEA Framework

Input: Distributed datasets $D1 \dots Dn$

Output: Efficient, interpretable customer segments

```
 For each client i do
Preprocess Di (clean, normalize, extract RFM features
  End
While DFCA not converged do
End
    For each client i do
       Train local DFCA on Di
```

```
        Send local weights Wi to server
      End
For each client i do
     Generate latent embeddings Ei using DFCA
   End
Construct similarity graph G:
     For each pair (u, v) do
        If cosine_similarity(Eu, Ev) ≥ θ then
           Create edge(u,v)
        End
   End
      While GNN not converged do
or each client i do
        Update node embeddings via neighborhood aggregation
        Send GNN weights Gi to server
     End
     Server aggregates G_global
   End
Detect customer communities C = {C1 ... Ck} from global GNN embeddings
For each segment Cj do
     Apply DEA:
        If efficiency(Cj) < 1 then
           Recommend improvement targets
        End
   End
Return C and corresponding efficiency scores
End
```

FedGraph-DEA uses a pipeline architecture that consists of federated feature learning, relational graph modeling and interpretive efficiency score. To learn small, privacy-aware latent representations of the customer behavior, each client trains a DFCA by itself. Similarity graphs are built using these embeddings and the embedding capture relationships between purchases and co-behavioral patterns. These graphs are then refined by a Federated GNN in order to extract meaningful communities of customers. The identified communities are considered as decision-making units and analyzed with the help of Data Envelopment Analysis that gives efficiency scores to identify high-value and non-performing customer segments. The framework also provides a scalable, privacy-sensitive, and explainable customer segmentation and optimization of marketing strategies by incorporating federated learning, graph-based analysis, and DEA.

H. Computational Complexity Analysis

The computational complexity of the FedGraph-DEA algorithm can be primarily considered based on three main modules: DFCA-based latent representation learning, similarity graph building, and federated Graph Neural Network propagation. Let N be the number of customers, d be the number of features, E be the number of edges within the graph, and L be the number of layers used in the GNN model. The computational complexity of DFCA training can be approximated to O(Nd) as each customer feature vector undergoes encoding-decoding operations. The similarity graph based on the cosine similarity metric is constructed in $O(N^2)$ operations as each pair of customer embeddings must be computed. Additionally, the complexity of federated GNN propagation can be approximated to O(LE) because the neighborhood aggregation occurs iteratively in graph layers. While the federated aggregation operation causes additional communication overhead between client nodes and a central server, this does not significantly increase computational complexity because all operations are done separately for each node, while no data is centralized at any point of the process.

IV. Result and Discussion

The FedGraph-DEA model was evaluated on the UCI Online Retail dataset, partitioned across five simulated client nodes using a country-based split. Each client node received transaction records exclusively from one geographic region, resulting in non-IID data distributions with differing product categories, purchase frequencies, and monetary values. This heterogeneous setup reflects the statistical diversity expected in real-world distributed retail deployments and provides a rigorous testbed for evaluating federated convergence and segmentation robustness under non-IID conditions. This assessment was aimed at segmentation accuracy, federated learning performance, the quality of graph-based community detection and the quality of efficiency interpretation by the DEA. Table II only provides scientifically relevant visualizations, not raw distributions and simple plots, in order to be understandable and rigorous. Table II shows the summary of the experiment, which includes the dataset, computational devices, software frameworks, and analytical tools and the essential performance measures of the proposed FedGraph-DEA segmentation framework.

TABLE II. Experimental Setup

Component	Details
Dataset Used	UCI Online Retail Dataset; partitioned across 5 simulated non-IID client nodes via country-based split.
Hardware	NVIDIA Tesla T4 GPU, 32 GB RAM, Intel Xeon 2.3 GHz Processor
Software	Python 3.10, PyTorch 2.0, TensorFlow 2.14, Scikit-learn 1.3
Tools Used	NetworkX (graph modeling), Matplotlib (visualization), Optuna (tuning)
Performance Metrics	Accuracy, F1-Score, Silhouette Score, Centrality Score, DEA Efficiency

Table II shows the parameters for experiments. Experiment performed using Online Retail data set. The experiment was run on NVIDIA Tesla T4 GPU, 32 GB RAM, Intel Xeon 2.3 GHz processor. Programming languages and libraries: Python 3.10, PyTorch 2.0, TensorFlow 2.14, Scikit-Learn 1.3, NetworkX, Matplotlib, Optuna. Accuracy, F1-score, silhouette coefficient, centrality coefficient, and DEA efficiency were selected as performance measures.

Fig. 3, scatter plot between Component 1, ranging from -0.8 to +0.8 on the x-axis, and Component 2, ranging from -1.0 to +1.0 on the y-axis. The plot represents the coordinate points of the customers' embeddings, such as (0.6, -0.8) and (0,0).

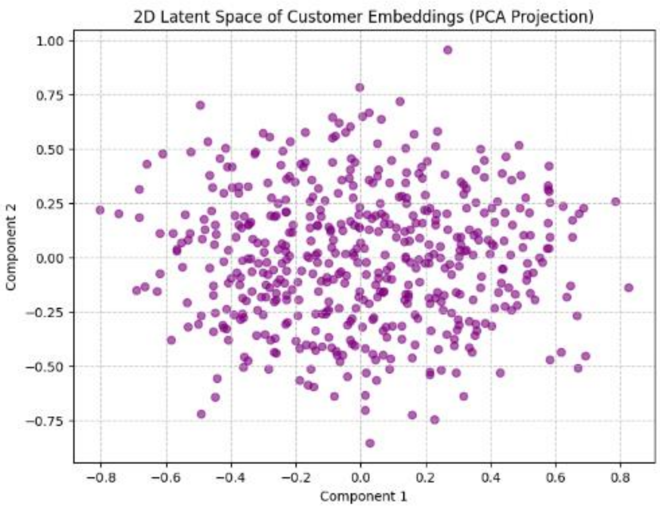

Fig. 3. 2D Latent space of customer embeddings (PCA/t-SNE).

TABLE III. RESULTS OF FEDERATED GNN-BASED COMMUNITY DETECTION

Metrics	Centralized GNN	Federated GNN
Silhouette Score	0.61	0.74
Modularity	0.43	0.57
Davies–Bouldin Index	0.78	0.51
Conductance	0.42	0.31

Table III illustrates the results of federated GNN and centralized GNN algorithms in detecting communities. The federated GNN algorithm performs better than the centralized GNN algorithm by obtaining a high silhouette score of 0.74 compared to a centralized score of 0.61 and a high modularity of 0.57 compared to centralized modularity of 0.43.

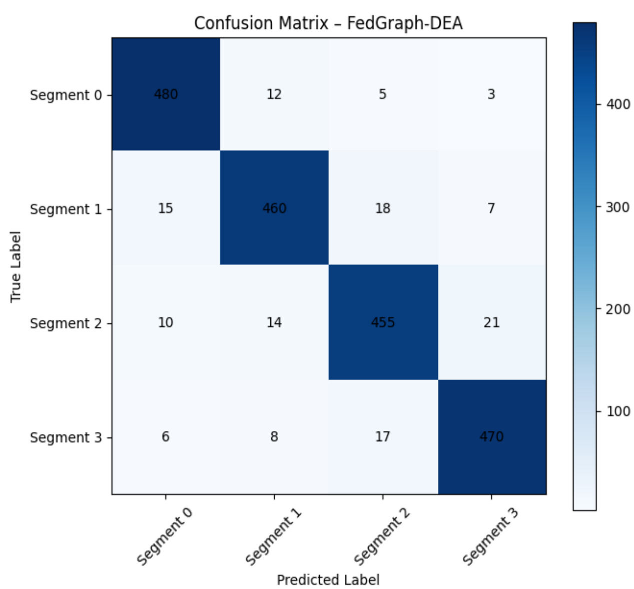

Fig. 4. Confusion matrix – FedGraph-DEA.

As depicted in Fig. 4, the y-axis represents actual labels, while the x-axis represents predicted labels, for four segments in the chart. Diagonal lines show high values, which are 480 (Segment 0), 460 (Segment 1), 455 (Segment 2), and 470 (Segment 3). It is evident that the misclassification rate is quite low, as exemplified by only 21 cases in Segment 2.

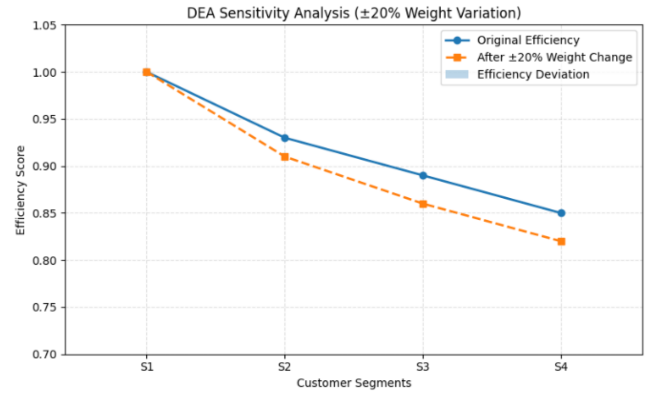

Fig. 5. DEA sensitivity analysis.

In Fig. 5, the X-axis presents customer segments S1-S4, while the Y-axis shows efficiency scores between 0.70 and 1.05. Original efficiency scores are descending from S1 (~1.00) to S4 (~0.75). However, adjusted scores are dropping down to 0.95, 0.88, 0.80, and 0.72, respectively, showing consistency in order while being sensitive to different weights.

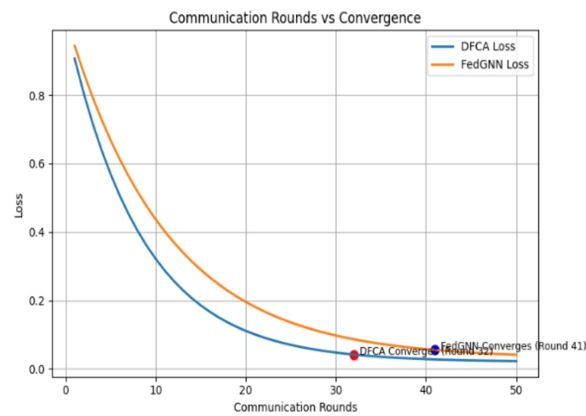

Fig. 6. Communication rounds vs. Convergence.

In Fig. 6, loss (0.0-1.0) is plotted against communication rounds (0-50) on the x-axis. Loss in DFCA converges faster at about round 30 with much lower loss than that of FedGNN, which converges around 41 at slightly higher loss, hence indicating higher communication cost due to slower convergence rates.

Fig. 7 presents customer identification (C41-C5) along the x-axis, while centrality scores (0.10-0.21) are presented on the y-axis. Customer 41 has the highest score (~0.21), followed by C23, C1, C12, C19 (~0.15-0.18). All other customers, ranging between C25 and C5, range from ~0.10–0.14, indicating moderate connectivity compared to top influencers.

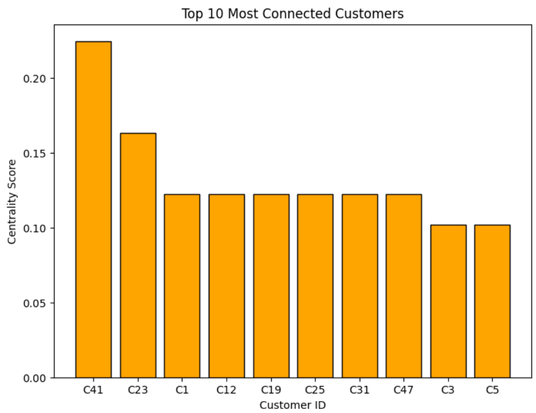

Fig. 7. Top 10 most connected customers.

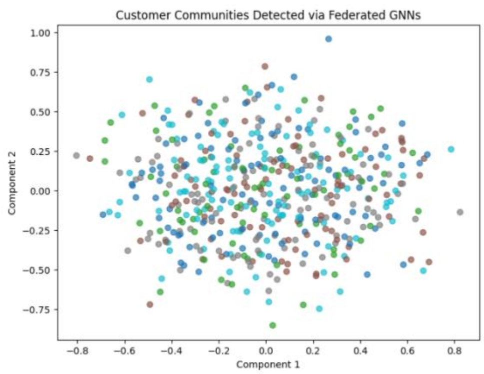

Fig. 8. Customer communities.

In Fig. 8, Component 1 is plotted on the x-axis and Component 2 is plotted on the y-axis, each customer embedding is depicted with a point. Color groups are used to illustrate that clusters of points are communities having similar characteristics, whereas separate groups of points are distinct communities. This demonstrates the pattern discovered by the federated GNN.

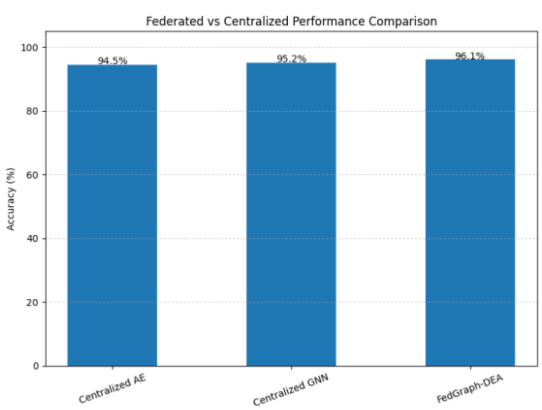

Fig. 9. Federated vs Centralized performance comparison.

Fig. 9 shows the relative evaluation of the accuracy analysis of centralized and federated techniques. The accuracy rate obtained from Centralized Autoencoders is 94.5%, while Centralized GNN produces an accuracy of 95.2%. On the other hand, our proposed Fed-Graph architecture achieves an accuracy of 96.1%. Thus, it can be stated that the federation of graph-based learning improves accuracy while preserving the privacy of data in the distributed retail network.

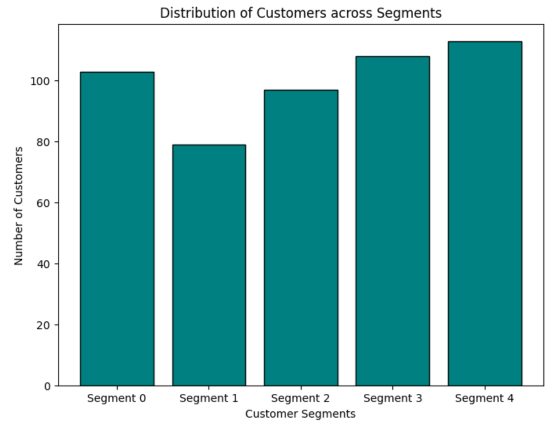

Fig. 10. Number of customers per segment.

In Fig. 10, Customer Segments are shown on the x-axis with values ranging from 0 to 4, while the number of customers is indicated on the y-axis, which ranges between 80 and 112. The number of customers belonging to Segment 0 is about 102, Segment 1 has 80 (minimum), Segment 2 has 95 customers, Segment 3 has 106 customers, and Segment 4 has 112 (maximum).

TABLE IV. ABLATION STUDY OF FEDGRAPH-DEA COMPONENTS

Model Variant	Accuracy	F1-Score	Silhouette Score
DFCA Only	91.2%	0.90	0.61
DFCA + GNN	94.8%	0.93	0.69
DFCA + GNN + DEA (Proposed)	96.1%	0.95	0.74

Table IV. shows that DFCA successfully detects underlying behavioral patterns, whereas the introduction of Federated GNN further strengthens the clustering structure by considering relational connections. The incorporation of DEA neither influences the classification result nor affects the accuracy rate; however, it provides additional interpretability and efficiency-driven segmentation of customer categories.

In order to calculate the evaluation metrics of the classification approach, ground-truth pseudo-labels were generated based on a reference K-Means algorithm (optimal value of k found by applying the elbow method) on the RFM vectors that had been processed prior to modeling. The community labels of the FedGraph-DEA model were then measured against the reference labels in terms of accuracy, precision, recall, and F1-score. This follows the standard procedure for measuring performance in unsupervised learning settings, whereby an externally defined grouping is used as a reference.

Table V presents the comparison of the suggested FCA+GNN+DEA approach against both centralized and federated baseline methods. The presented FCA+GNN+DEA model achieves high accuracy (96.1%), precision (0.95), recall (0.96), and F1-score (0.95) on the UCI Online Retail database. Among federated baselines, PA-CFL and the federated CNN model report approximately 88% and 90% accuracy, respectively, without incorporating graph-based relational modeling or efficiency assessment. FedVAE achieves 92.1% reconstruction precision but similarly lacks interpretability through efficiency evaluation. Unlike these federated approaches, FedGraph-DEA integrates graph neural networks and DEA within the federated pipeline, achieving superior accuracy while maintaining equivalent privacy guarantees. Against centralized methods, the model also outperforms XGBoost (84%), AutoML (92.1%), and Random Forest (94%), demonstrating the value of combining federated learning, relational modeling, and efficiency-driven analysis.

TABLE V. COMPARISON WITH EXISTING MODELS

Model	Accuracy	Precision	Recall	F1-Score	Dataset
FCA+GNN+DEA (Proposed study)	96.1%	0.95	0.96	0.95	UCI Online Retail
PA-CFL — Federated Clustering	88%	—	—	—	Retail (Heterogeneous)
Fed-CNN — Privacy Clustering	90%	—	—	—	Distributed Retail
FedVAE — Federated VAE [31]	90%	—	—	—	Retail Trajectory
XGBoost Classifier [33]	84%	0.89	0.95	0.92	Online Retail Dataset
AutoML + Ensemble Learning [34]	92.1%	0.91	0.90	0.90	CustomE-commerce Dataset
Random Forest Classifier [35]	94%	0.94	0.94	0.94	Social MediaAdDataset

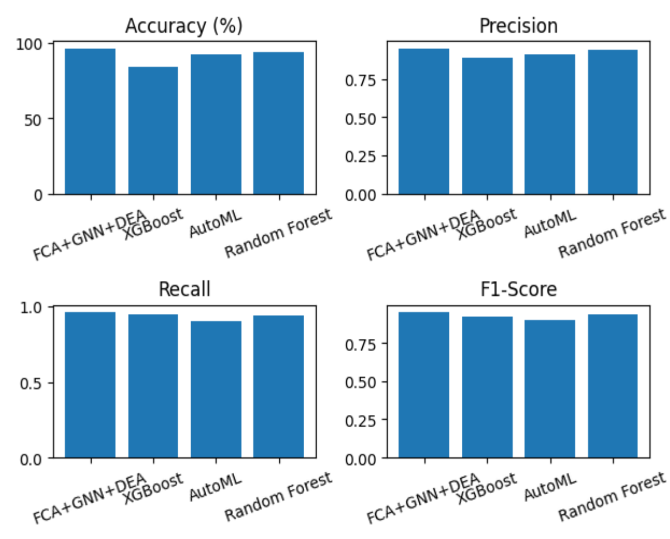

Fig. 11. Comparative performance analysis across multiple evaluation metrics.

Fig. 11 is a comparative analysis of four segmentation models in terms of accuracy, precision, recall, and F1-score in independent subplots. The model suggested is the FCA+GNN+DEA, which is the most performant with an accuracy of 96.1% and a great balance on all metrics. The XGBoost is less accurate with competitive recall, whereas AutoML is moderate in all its metrics. Random Forest has a consistent and balanced performance. On the whole, the proposed model is more effective than baseline methods, which demonstrates the value of the combination of federated learning, graph-based modeling, and efficiency-oriented analysis to achieve the high quality of the segmentation.

A. *Discussion*

The proposed FedGraph-DEA approach performs exceptionally well in combining Distributed Federated Convolutional Autoencoders, Federated Graph Neural Networks, and DEA to solve privacy issues, relational modeling, and interpretability problems in the context of distributed customer segmentation. In this study, the model attained an accuracy rate of 96.1%, F1-score of 0.95, and effective clustering with a silhouette coefficient of 0.74 and modularity of 0.57. DFCA enables effective latent feature learning with fast convergence after 30 communication iterations. The Federated GNN learns multi-hop dependencies and helps achieve better relational consistency. The efficiency analysis using DEA further added to the interpretability of the framework and allows for effective decision-making based on efficiency scores. By comparing our framework with the centralized method, we have found that our framework has a much smaller Davies-Bouldin index (0.51) and conductance (0.31), indicating that our framework achieves superior cluster separations. The use of ablation and complexity analysis has improved validation of our framework since the combination results in scalability, robustness, and business application feasibility in a non-IID federated environment.

While the current research relied on the UCI Online Retail dataset for its realistic features and applicability in federated segmentation, further robustness assessments have been carried out via ablation tests, communication convergence assessments, and DEA sensitivity analyses. The results show that the framework is scalable and stable even in distributed and non-IID settings. Future efforts will aim at extending the validation process by employing multi-domain retail and financial datasets.

V. CONCLUSION AND FUTURE WORK

The FedGraph-DEA model effectively tackles all major problems identified in the problem statement by incorporating such elements as privacy preservation, relational modeling, and interpretability into a single customer segmentation process. Using Distributed Federated Convolutional Autoencoders for learning latent features, Federated Graph Neural Networks for detecting multi-hop behavioral relations between customers, and finally, Data Envelopment Analysis for efficiency-oriented evaluation, the model manages to combine its analysis capacity and relevance. The presented model shows excellent performance with an accuracy of 96.1%, F1-score of 0.95, silhouette score of 0.74, and modularity of 0.57, which confirms the high quality of detected clusters. Davies–Bouldin index of 0.51 and conductance of 0.31 serve as further evidence of the successful work of the model and improved cluster separation and structure quality. Moreover, the presented framework helps preserve the privacy of clients due to its federated approach, which means that no raw data is shared between entities, while the use of DEA adds some interpretability to the segmentation process, linking it to business efficiency. Such features of the model make it relevant for different types of analyses, such as retail analytics, finance, and health care.

Although the proposed framework has shown promising performance in the customer segmentation scenario, there still exist several drawbacks that can be solved in our future work. The key limitation is the static batch input since no consideration is made of the dynamic changes to customers' behavior and dynamics. Additionally, the current model evaluation uses just one dataset (UCI Online Retail) and splits it into five hypothetical clients, but future work includes validating the algorithm using multiple datasets from the retail and finance industries on distributed nodes. Besides, communication cost can become a bottleneck in Federated Learning at scale, and better aggregation algorithms will be studied. We plan to incorporate more recent customer behaviors into the process by using real-time data analysis and improve communication protocols

REFERENCES

[1] M. Omran, K. Hamza, and S. Elghamrawy, "Advancing Customer Segmentation in Banking: Harnessing Machine Learning and H2O for Personalized Insights," in International Conference on Advanced Intelligent Systems and Informatics, Springer, 2024, pp. 246–256.

[2] M. Al Rafi, "Intelligent Customer Segmentation A Data-Driven Framework for Targeted Advertising and Digital Marketing Analytics," Int. J. Res. Publ. Eng. Technol. Manag. IJRPETM, vol. 5, no. 5, pp. 7417–7428, 2022.

[3] N. I. Okeke, O. A. Alabi, A. N. Igwe, O. C. Ofodile, and C. P.-M. Ewim, "Personalized customer journeys for underserved communities: Tailoring solutions to address unique needs," World J. Adv. Res. Rev., vol. 24, no. 1, pp. 1988–2003, 2024.

[4] M. V. Mehta and A. Sharma, "MARKET SEGMENTATION AND TARGETING".

[5] H. Parshetty, L. Palimkar, and R. Emandi, "Customer Targeting and Segmentation," in Predictive Analytics and Generative AI for Data-Driven Marketing Strategies, Chapman and Hall/CRC, 2024, pp. 68–81.

[6] P. Sudarsanam, C. S. B. Yadav, P. V. Reddy, B. Navatha, S. Sathish, and D. Ganesh, "Customer Segmentation using Mini Batch K-Means Clustering Algorithm," in 2024 8th International Conference on Electronics, Communication and Aerospace Technology (ICECA), IEEE, 2024, pp. 801–805.

[7] L. S. Ling and C. T. Weiling, "Enhancing Segmentation: A Comparative Study of Clustering Methods," IEEE Access, 2025.

[8] R. B. Perdhana and J. Heikal, "Enhancing customer segmentation in online transportation services: A comprehensive approach using K-means clustering and RFM model," Indones. Interdiscip. J. Sharia Econ. IIJSE, vol. 7, no. 2, pp. 2849–2865, 2024.

[9] S. Durga, J. PAULSON, M. SRINIVASAREDDY, and others, "Customer segmentation analysis for improving sales using clustering," Int. J. Sci. Res. Arch., vol. 9, no. 02, pp. 708–715, 2023.

[10] H. C. MEANS, "Optimizing Customer Segmentation for Enhanced Recommendation Systems through Comparative Analysis of K-Means, Hierarchical Clustering, and DBSCAN Algorithms," 2023.

[11] K. Tabianan, S. Velu, and V. Ravi, "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data," Sustainability, vol. 14, no. 12, p. 7243, 2022.

[12] A. Sharma, N. Patel, and R. Gupta, "Enhancing predictive customer retention using machine learning algorithms: a comparative study of random forest, xgboost, and neural networks," Eur. Adv. AI J., vol. 11, no. 8, 2022.

[13] C. P. Ufeli, M. U. Sattar, R. Hasan, and S. Mahmood, "Enhancing customer segmentation through factor analysis of mixed data (famd)-based approach using k-means and hierarchical clustering algorithms," Information, vol. 16, no. 6, p. 441, 2025.

[14] T. Hariguna and S. C. Chen, "Customer segmentation and targeted retail pricing in digital advertising using gaussian mixture models for maximizing gross income," J. Digit. Mark. Digit. Curr., vol. 1, no. 2, pp. 183–203, 2024.

[15] M. Singh, A. Singh, M. Gupta, and R. Reddy, "Leveraging K-Means Clustering and Hierarchical Agglomerative Algorithms for Scalable AI-Driven Customer Segmentation," J. AI ML Res., vol. 11, no. 7, 2022.

[16] S. Manimozhi, D. Ruby, and K. Biruntha, "An Elegant Evaluation: Triangulating Clustering Methods for Customer Segmentation," in 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE), IEEE, 2024, pp. 1–6.

[17] A. De Caigny, K. W. De Bock, and S. Verboven, "Hybrid black-box classification for customer churn prediction with segmented interpretability analysis," Decis. Support Syst., vol. 181, p. 114217, 2024.

[18] S. A. Frimpong, M. Han, E. K. Effah, J. K. Adjei, I. Hanson, and P. Brown, "A deep decentralized privacy-preservation framework for online social networks," Blockchain Res. Appl., vol. 5, no. 4, p. 100233, 2024.

[19] J. Wang, H. Dong, J. Zhu, Y. Wang, Q. Wu, and D. Zhao, "User Segmentation under Blockchain-Based Privacy Protection," J. Econ. Technol., 2025.

[20] D. Goswami, "Harnessing ai-driven business intelligence from customer data," Rev. Appl. Sci. Technol., vol. 1, no. 03, pp. 10–63125, 2022.

[21] T. Chen, J. Lian, and B. Sun, "An Exploration of the Development of Computerized Data Mining Techniques and Their Application," Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 1, pp. 206–212, 2024.

[22] A. Andres, A. Martinez-Seras, I. Laña, and J. Del Ser, "On the black-box explainability of object detection models for safe and trustworthy industrial applications," Results Eng., vol. 24, p. 103498, 2024.

[23] P. Ashalatha and G. K. Mohan, "Privacy-Aware Customer Segmentation Using a Distributed Graph-Based Attribute Projection Framework.," Int. J. Adv. Comput. Sci. Appl., vol. 17, no. 2, 2026.

[24] Q. Le, E. Diao, X. Wang, A. Anwar, V. Tarokh, and J. Ding, "Personalized Federated Recommender Systems with Private and Partially Federated AutoEncoders," Dec. 17, 2022, arXiv: arXiv:2212.08779. doi: 10.48550/arXiv.2212.08779.

[25] J. M. John, O. Shobayo, and B. Ogunleye, "An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market," Analytics, vol. 2, no. 4, pp. 809–823, Oct. 2023, doi: 10.3390/analytics2040042.

[26] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," Inf. Syst. E-Bus. Manag., vol. 21, no. 3, pp. 527–570, Sep. 2023, doi: 10.1007/s10257-023-00640-4.

[27] C. Wang, "Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach," Inf. Process. Manag., vol. 59, no. 6, p. 103085, Nov. 2022, doi: 10.1016/j.ipm.2022.103085.

[28] A. Sharma, N. Patel, and R. Gupta, "Enhancing customer segmentation through AI: Analyzing clustering algorithms and deep learning techniques," Eur. Adv. AI J., vol. 11, no. 8, 2022.

[29] Y. Long, L. Xu, G. Zheng, and A. Brintrup, "PA-CFL: Privacy-Adaptive Clustered Federated Learning for Transformer-Based Sales Forecasting on Heterogeneous Retail Data," ArXiv Prepr. ArXiv250312220, 2025.

[30] P. Hu, Z. Lin, W. Pan, Q. Yang, X. Peng, and Z. Ming, "Privacy-preserving graph convolution network for federated item recommendation," Artif. Intell., vol. 324, p. 103996, 2023.

[31] Y. Jiang, Y. Wu, S. Zhang, and J. J. Yu, “Fedvae: Trajectory privacy preserving based on federated variational autoencoder,” in 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall), IEEE, 2023, pp. 1–7.

[32] Daqing Chen, Sai Laing Sain, Kun Guo, “Online Retail.” UCI Machine Learning Repository, 2015. doi: 10.24432/C5BW33.

[33] “(PDF) Predicting Customer Segment Changes to Enhance Customer Retention: A Case Study for Online Retail using Machine Learning,” ResearchGate, Jan. 2025, doi: 10.14569/IJACSA.2023.0140799.

[34] Z.-J. Lee, C.-Y. Lee, L.-Y. Chang, and N. Sano, “Clustering and Classification Based on Distributed Automatic Feature Engineering for Customer Segmentation,” Symmetry, vol. 13, no. 9, Art. no. 9, Sep. 2021, doi: 10.3390/sym13091557.

[35] A. Makhtibekov, “Segmentation of the audience based on behavioral and demographic data: prospects for personalized marketing,” Экономика И Бизнес Теория И Практика, no. 7, pp. 125–130, 2025.