Privacy-Aware Customer Segmentation Using a Distributed Graph-Based Attribute Projection 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 vital role in Business Intelligence (BI) by enabling organizations to understand customer behavior, enhance personalization, and support informed decision-making. Conventional segmentation approaches, including K-Means clustering, hierarchical methods, and hybrid deep learning models, often face limitations when handling high-dimensional customer data and typically lack built-in mechanisms to address privacy concerns. As customer analytics increasingly relies on sensitive personal information, these limitations pose significant challenges for responsible data-driven applications. To overcome these issues, this study introduces a Distributed Graph-Based Attribute Projection Framework (GAPF) for privacy-aware customer segmentation. The key novelty of the proposed framework lies in its ability to minimize sensitive attribute exposure while preserving meaningful relational patterns among customers through graph-based representations. GAPF employs a distributed processing pipeline that integrates attribute projection to reduce identifiability, heuristic-driven customer similarity graph construction, graph convolutional network (GCN)–based feature learning, and community detection for final segmentation. The framework is implemented using Python, NetworkX, and PyTorch Geometric and evaluated on the Mall Customers dataset and large-scale anonymized synthetic data to assess scalability. Experimental results demonstrate that GAPF achieves superior segmentation performance, with an accuracy of 98%, precision of 92.5%, recall of 94.0%, and an F1-score of 93.2%, while also exhibiting efficient execution and reduced privacy risk. These findings confirm GAPF as a robust and practical solution for privacy-aware BI applications.**

Keywords—Business intelligence; privacy-aware customer segmentation; graph-based learning; federated analytics; graph neural networks

I. Introduction

Over the past years, customer segmentation has been integrated into the BI systems as one of the foundational aspects that facilitate organizations to study customer behavior, discover market trends, and provide customized products and services [1]. As online platforms are becoming increasingly popular, businesses can access large volumes of customer data in a variety of places, such as transaction records, online interactions, mobile apps, and loyalty programs [2], [3]. Successful segmentation of such data helps in strategic decision-making in the areas of retail, finance, telecommunication and e-commerce [4], [5]. Nevertheless, the growing magnitude, size and delicacy of customer information pose a serious obstacle to conventional analytical practices [6]. Traditional segmentation methods, such as K-Means clustering, hierarchical clustering, and density-based clustering, mostly focus on distance-based similarity calculations in high-dimensional feature spaces. Although these methods are computationally efficient, they may lack scalability, robustness to noise, and the ability to identify non-linear relationships between customers. [7]. Although these techniques are computationally effective, they are frequently characterized by low scalability, noise sensitivity and low power to describe non-linear and intricate associations among customers [8]. Recent developments in hybrid deep learning-based segmentation methods have increased the representational power, but they generally need centralized access to data and large-scale feature exposure [9]. Consequently, these methods have grave privacy concerns, particularly when handling personally identifiable information (PII) and sensitive behavioral characteristics [10], [11]. Due to the rising regulatory limitations like the General Data Protection Regulation (GDPR) and the rising awareness of data abuse, privacy preservation has become a paramount need in contemporary BI systems [12], [13]. Current privacy-saving methods, such as data anonymization, k-anonymity, and differential privacy, tend to impair the quality of analysis, or can interfere with relational patterns that can be used to perform accurate segmentation [14] [15]. Besides, the vast majority of the current segmentation models consider privacy as an after-processing, but not an in-processing, component of the modeling pipeline, resulting in suboptimal trade-offs between data utility and privacy protection [16].

Graph-based learning has been proposed as an effective paradigm to overcome the drawbacks of these methods in order to model complex relations in structured data. The similarities and interactions among customers are naturally represented in graph forms, giving more expressive and contextual segmentation. Nevertheless, the existing graph-based customer analytics systems do not pay much attention to privacy-conscious attribute processing and do not scale to distributed BI systems. This is the gap that demonstrates the necessity of an integrated framework that would incorporate both graph learning and privacy-conscious data transformations in a scalable way. In this regard, a Privacy-Aware Customer Segmentation framework is proposed in this study based on a Distributed GAPF. The proposed solution proposes attribute projection schemes to minimize attribute identifiability and maintain critical relational systems among customers. The framework facilitates the process of robust segmentation by using community detection in the context of a similarity graph constructed on the projected attributes and representation learning with the help of GCNs. Scalability of large volumes of data is an additional benefit with the distributed design of GAPF that does not affect privacy limits. The overall contributions of this work include three folds: (1) the creation of a new graph-based attribute projection strategy, where the privacy awareness is incorporated directly into the segmentation pipeline, (2) the development of a scalable and distributed graph learning framework, used to carry out customer segmentation, and (3) a comprehensive experimental analysis, showing that the given approach outperforms the traditional techniques of clustering and hybrid learning methods in terms of performance, scalability, and low privacy risk. On the whole, the framework suggested is a sensible and accountable customer segmentation solution to the contemporary BI framework, which balances the efficiency of the analytics with the privacy of information.

A. Problem Statement

Although customer segmentation has been used extensively in Business Intelligence systems, the current methods of segmentation, like traditional clustering and hybrid deep learning models, have serious constraints on large-scale, high-dimensional, and sensitive customer information [17]. They are usually based on centralized processing and access to raw attributes, which exposes them to privacy risks and non-adherence to regulations [18]. Moreover, privacy protection methods that are more traditional are also likely to decrease the accuracy of analytical methods due to data distortion or the incomplete preservation of meaningful relational patterns between customers [19]. Consequently, scalable segmentation schemes can lack scalable segmentation frameworks that can successfully trade segmentation efficiency and privacy consciousness [20]. This brings a pressing requirement for an integrated solution that can combine privacy-preserving attribute manipulation with relational learning in a distributed and scalable form, and at the same time be of high segmentation quality as in the case of real-world Business Intelligence applications.

B. Research Motivation

This study is motivated by the fact that organizations are becoming more dependent on customer data to make intelligent business decisions and the rising concerns about data privacy, data security and adherence to regulations. Even though sophisticated analytics and learning models have improved the accuracy of segmentation, they may need complete access to sensitive customer properties, which expose organizations to privacy risks and restrict their application in practice. Simultaneously, current privacy-preserving methods often reduce the analytical value or are not able to reflect on intricate customer associations. The necessity of such challenges is the need to have a segmentation framework, which can maintain the privacy of the customers, at the same time, model relation patterns efficiently, and scale to large datasets. Through the combination of privacy-conscious attribute projection and graph learning in a distributed environment, the study will address the gap between the responsible data usage and high-performance customer analytics in contemporary Business Intelligence systems.

C. Significance of the Study

The study is important because it fills an important gap in modern Business Intelligence by incorporating the conservation of privacy directly in the customer segmentation process without affecting the performance of the analysis. The proposed graph-based attribute projection framework incorporates privacy awareness in data representation and learning, unlike the traditional approaches to segmentation, which regard privacy as an extrinsic limitation, and allows taking responsible care about sensitive customer data. The research effort provides a scalable and distributed segmentation model that is able to process high-dimensional and large-scale data and retain relational structure that is vital in performing an accurate customer profiling process. In practical terms, the framework contributes to adherence to data protection regulations and improves trust in the data-driven decision systems, which makes it very relevant to practical application in various areas, including retail, finance, and telecommunications.

D. Key Contribution

Introduced a new privacy-aware attribute projection algorithm that minimized sensitive attribute identifiability and still maintained valuable customer relations.

Developed a distributed graph-based customer segmentation framework: Built a scalable framework with high-dimensional data and efficient scale to large-scale data.

Community detection based on an integrated Graph Convolutional Network (GCN) representation learning to enhance segmentation accuracy and robustness.

Demonstrated to outperform conventional clustering and hybrid learning methods in a large scale of experimental analysis regarding accuracy, scale, and reduced privacy risk.

E. Rest of the Sections of the Study

The following is the rest of the section: related works are given in Section II. The suggested methodology framework is presented in detail in Section III. The results of the study are shown in Section IV. Lastly, Section V wraps up the conclusions and future of the suggested framework.

II. Related Works

Khan et al. [21] focus on the use of the BAT-ANN model for the behavior of customer churn in the telecom sector. Based on the Big ML dataset, the model establishes 89.2% accuracy, which reveals the ability of the model to pinpoint significant factors that lead to customer churn. The BAT-ANN model is useful for companies which seek to develop specific retention initiatives and thus improve the overall customer goodwill and revenue. The study contributes a framework for churn prediction by utilizing the Bat Algorithm for feature optimization and an ANN for classification. However, although the proposed model displays good performance, it is not very effective in more real-world environments, for instance, working with streaming data.

This limitation is because of the computational dimension of ANN and the difficulty of processing continuously changing patterns of data. Furthermore, the permutation of the model across different industries is unknown owing to the great difference in customer behavior and churn factors across domains. While this study has a good start in providing an understanding of churn prediction models to help in their improvement, the work could endeavor to build on the scalability and robustness of the models in order to address the dynamic and complex nature of various real-world settings.

Lu [22] investigates the problem of attaining utility and privacy simultaneously in social media data mining by using a pairwise GCN. The model aims at providing considerable levels of protection against adversarial threats, Still delivers stupendous degree of accuracy so as to effectively mine data with high security standards. In turn, to enhance privacy protection during the training of the models, federated learning is incorporated into the learning process. They do this by maintaining data decentralization, so the data stays local to the owner's environment, thus eliminating privacy concerns while at the same time facilitating collaborative model training. The experiments carried out in this study reveal that the proposed approach is realistic by indicating how privacy and utility can reasonably be traded off. However, the study finds the following challenges: Therefore, one main workload is the scalability of the proposed federated learning framework, especially in big social media applications that demand more computational and communication power. In addition, it can be seen that the model is good and performs well on a test set, though it might be less effective in application on real-world data with various, noisy, and missing data. Such conditions, typical in social media data, can pose a limit on how well the model generalizes. There are, however, several challenges to this approach that require further study to confirm the obtained results and further expand the methodological applicability in similar conditions and types of SNS, such as mixing with other user attributes, and the dynamic nature of SM features.

Li et al. [23] provide a systematic review of privacy preservation schemes in MEC task offloading. The study categorizes privacy-preserving methods into three main areas: data storage, transmission security, and choosing the appropriate place for data storage. Each category is accompanied by various techniques that are explained in detail in their mechanism of data privacy while offloading. The study focuses on the privacy requirements in MEC environments where different data may be, most of the time, highly confidential and have to be processed and transmitted through distributed networks. Nevertheless, the given study has several significant limitations. Comparatively, it does not feature quantitative assessments or accuracy indices for the aforesaid techniques, and hence there is no real measure of the effectiveness of them all. Altogether, the effectiveness of the proposed methods in similar conditions is shown, but their applicability for real-world situations is not explored in detail. More complex edge environments with different and changing privacy needs (for instance, industrial or healthcare scenarios) might be problematic in this regard. The study suggests that work should concentrate more on concrete applications and strategies for overcoming the conflict of interest between privacy, performance, and scalability. I believe that more such efforts are necessary in order to close the gap between the ideas explored by study and the actual implementation of privacy-preserving MEC solutions.

Anand and Lee [17] explore the feasibility of GANs for achieving customer data privacy with the retention of critical data features. The benchmarks of tradeoff management are improved by their proposed model while still keeping a balance between accuracy and privacy. Because the approach produces simulated data that presents similar features to real data but does not include any personally identifiable information, it is helpful when used in practical applications, including setting artificial price marks and marketing strategies to target shoppers. Thus, the given study illustrates the applicability of the proposed model for the analysis of big real-time customer data in industries. Nevertheless, it has also pointed of weakness. Still, there are some drawbacks connected with the computational heuristic nature, especially if the organization has problems with its computational power. Moreover, there is always a risk of overfitting that may affect the model's performance in different contexts and marketing datasets. This issue may cause a restriction on applying this method in different environments with different customer behavior. Despite the study's contributions to the utilization of the method for privacy conservation, more study needs to be conducted to enhance the computational complexity, and minimize overfitting and improve the generality of the model.

Sharma, Patel, and Gupta [18] propose a framework integrating K-Means clustering with Deep Learning to improve the customer segmentation process. To address the above issue, the authors combine the advantages of K-Means and deep learning while proposing a new approach that features a higher level of segmentation detail and better accuracy. When applied to the heteroscedastic labeling problem in large-scale retail datasets, the proposed model has shown substantial promise in modeling customer behavior and thus can be useful in determining efficient marketing strategies. Due to the employment of deep learning, the method is capable of analyzing intricate structures in the customer data so that various business campaigns can be more effectively developed. However, the hybrid model does have its drawbacks, which will be discussed in this article. Furthermore, the optimization of the model is also extremely sensitive to the quality of the clustering achieved by K-Means. While clustering, if done in an improper way, affects the result very badly and also reduces the efficiency of the whole process. Consequently, although the described hybrid model can be considered an improvement over the classical approaches to segmentation, additional study is required to address these issues. Possible future work on this matter could be oriented towards minimizing the computational complexity as well as enhancing the stability of the initial clustering phase, which makes the suggested model applicable for various businesses and datasets.

Ali et al. [19] investigated the feasibility of using federated learning for predictive maintenance in Industry. FL ensures that participants can train datasets independently without sharing them with different parties and, in the process, allows organizations to leverage some collective conclusions while protecting individual data. In the course of the study, several examples illustrating the effectiveness of FL for improving maintenance performance while preserving data integrity are also presented. As the study shows, there are quite a few drawbacks that should be solved before the implementation of such a system on a mass scale. There are several challenges, such as data heterogeneity, generated by the variation in data format and data quality of devices. Yet another factor—a communication overhead—is a problem here, as multiple interactions between devices and servers might slow down the process. However, the necessity of reliable privacy-preserving methods is most important in large industrial applications. This holds especially because the study identifies some challenges in using FL for predictive maintenance and the need to overcome these challenges to maximize the benefit of FL in achieving the need. Further should be targeted towards FL systems that would be manageable and applied successfully to the complex and variable necessities in industries.

From the study made by Tabianan, Velu, and Ravi [20], it is clear that customer segmentation acts as a key factor in increasing the profitability of e-commerce systems. Customer segmentation is a process that aims at categorizing customers to enhance the efficiency of satisfying the needs of each category of customers. The proposed method seeks to optimize both within-cluster homoscedasticity and between-cluster heteroscedasticity so that customers belonging to a given cluster exhibit similar behavior to the other customers in the same cluster but different from those in other clusters. Such a segmentation puts a business in a better position to sell its products to different customers through promotional tools, services, and marketing strategies, all of which increase customer satisfaction, loyalty, and profitability. However, the application of the current study has some drawbacks because of the reliance on the adopted K-Means algorithm. The problem with the first few choices affects the degree of cluster reliability and accuracy of a segment. Moreover, real life often embodies different behavior of customers and different patterns which shift with time, and this is another factor which opposes K-Means, which is not capable of handling these issues. It also works poorly with datasets containing outliers or datasets where the density of clusters varies a lot, which may be true in some more complex e-commerce environments. These issues bring about the necessity of developing new and more flexible clustering techniques to provide e-commerce systems with more stability in the face of the constantly changing customer behavior.

Ebrahimi et al. [24] conducted an experiment to investigate the relationship between Social Network Marketing (SNM) and Consumer Purchase Behavior (CPB) within the context of Facebook Marketplace. To reveal specific consumer segments, the empirical study used two types of Structural Equation Modeling (SEM) and two machine learning methods that are not supervised: hierarchical clustering and K-Means. The online sample of Facebook Marketplace users in Hungary was taken and provided data, yielding 475 participants and a high response rate of 98.1. Following the screening, there were 466 valid responses that were left to be analyzed. The use of clustering of the consumers identified nine consumer groups in terms of both demographic and behavioral attributes. This division allowed agencies to create specific marketing plans for each group, and this enhanced customer engagement and sales. Nevertheless, the external validity of the results can be considered rather low because the research was conducted in one social networking site and a particular geographic area. The biggest methodological weakness is the use of static clustering algorithms that make the assumption of the stability of consumer behavior. Since the interactions among the users on social networks are dynamic and undergo constant changes, the fixed segmentation models might not reflect the real-time changes in the preferences of the consumers. Based on this, the authors emphasize that more sophisticated, dynamic clustering methods are required that will be able to capture the shifts in behavioral patterns, and more precise and responsive marketing strategies in the fast-moving digital market environment are guaranteed.

TABLE I. SUMMARY OF RELATED WORKS

Author(s)	Techniques	Advantages	Disadvantages
Khan et al. [21]	BAT-ANN (Bat Algorithm + ANN)	Achieved 89.2% accuracy on churn prediction and identified key factors for customer churn.	Not effective for streaming/real-time data, and poor generalizability across industries.
Lu [22]	Pairwise GCN + Federated Learning	Preserves privacy with decentralized training and high accuracy under security constraints.	Scalability issues in large social media applications are sensitive to noisy/missing data
Li et al. [23]	Privacy-preserving schemes in MEC	Comprehensive review of MEC privacy techniques, and identifies key mechanisms for storage, transmission.	Lacks quantitative effectiveness measures, and no solutions for dynamic environments.
Anand & Lee [17]	GANs for privacy-preserving data	Balances accuracy and privacy, and generates synthetic but realistic data	High computational cost and risk of overfitting.
Sharma et al. [18]	K-Means + Deep Learning (Hybrid)	Captures intricate customer patterns, and improves segmentation detail & accuracy	High computational cost, and requires robust infrastructure.
Ali et al. [19]	Federated Learning for maintenance	Preserves data privacy by local training, and effective in industrial settings.	Data heterogeneity challenges, and communication overhead
Tabianan et al. [20],	K-Means for behavioral segmentation	Groups customers based on behavior, product & events Improves satisfaction, loyalty, and profitability.	Sensitive to initial cluster centers, and poor performance on dynamic/outlier-rich datasets.- Cannot adapt to evolving behaviors.
Ebrahimi et al. [24]	SEM + Hierarchical & K-Means	Identifies key SNM factors influencing purchase behavior, and defines customer clusters effectively.	Sample bias due to convenience sampling, and Limited generalizability to other platforms.

Table I presents a compact evaluation of the available methods in customer grouping and privacy protection. It points out the methods used, their advantages in regard to increasing accuracy of segmentation, confidentiality, scalability, or significant disadvantages like computational complexity, scalability, or lack of applicability. This comparative analysis emphasizes on the issues of robust, privacy-sensitive and sizeable frameworks.

III. Privacy-Aware Customer Segmentation Using Graph-Based Attribute Projection

The proposed methodology is rooted in a privacy-aware and decentralized graph-based segmentation pipeline that is aimed at creating a balance between analytical and data protection. The framework combines preprocessing, attribute projection that is privacy sensitive, distributed graph construction, and graph neural representation learning in a single architecture that can be used in scalable Business Intelligence applications. A similarity measure based on a heuristic is employed to build a customer graph, which reflects the pattern of relationships among people based on the projected attributes. GCNs are then used to learn rich representations of customers using the combined attributes information and structural information. Lastly, on the learned graph embedding, community detectors are applied to create meaningful and privacy-preserving customer segments that can be used in Business Intelligence applications at scale. The distributed execution of the suggested framework was implemented with the help of PyTorch Distributed Data Parallel (DDP), which allows parallel representation learning in the form of a graph that is distributed among several computational units. This decentralized structure enhances training effectiveness and facilitates scalability by operating with more customers in Business Intelligence applications. Fig. 1 shows the proposed methodology workflow.

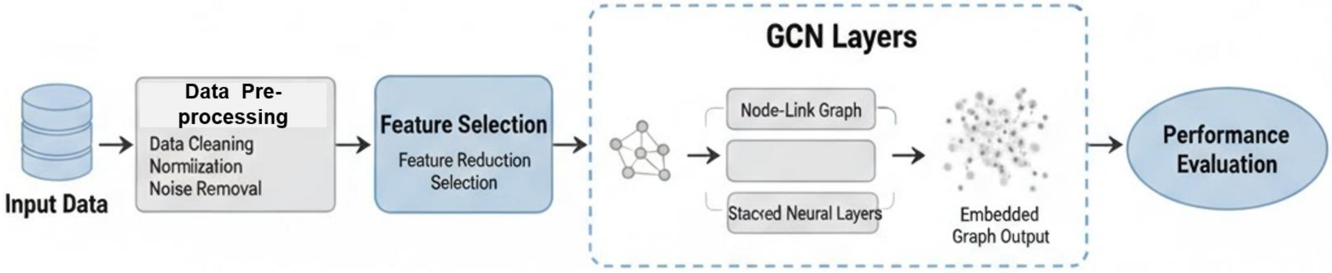

Fig. 1. Proposed model workflow.

A. Dataset Description

A famous dataset is the Mall Customers data set availed by Kaggle because it is easy and applicable to the customer segmentation study (Table II) [25]. It includes 200 single customer records, the properties will be CustomerID, Gender, Age, Annual Income (k), and Spending Score (1-100), all of which will be required to project the trend of demographics and behavior. In this study, GAPF has been used on the data. Attributes are seen as nodes, and relationships are drawn among the similar customers in terms of attributes, be it demographic or spending habits. The graphical display facilitates customer privacy-friendly clustering and retention of the relational structures to support scalable customer segmentation and BI applications. Although the Mall Customers dataset is very small, it can be utilized in reference in testing segmentation frameworks. Scalability was also achieved by having the data expand at runtime under control.

TABLE II. Sample of the Mall Customers Dataset

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1–100)
0001	Male	19	15	39
0002	Male	21	15	81
0003	Female	20	16	6
0004	Female	23	16	77
0005	Female	31	17	40

Despite the fact that Mall Customers' data is relatively small and it is used mostly as a demonstration of the concept, the data offers an appropriate measure of the segmentation frameworks in the validity testing. Scalability was experimented with by expanding the data under control in the real time. Nonetheless, confirmation in larger real-world data sets is required to substantially ensure the effectiveness of the framework in a large-scale Business Intelligence setting.

B. Data Preprocessing

The first step in data mining is an essential step of data preprocessing that guarantees the quality, uniformity, and appropriateness of the customer data in terms of graph-based learning and privacy-aware segmentation. Raw customer data are often filled with missing values, noise, non-homogeneous scales and high-dimensional attributes that negatively impact similarity computation and subsequent graph modeling. Thus, a system preprocessing pipeline is employed before attribute projection and graph building.

1) Handling missing values: The absence of attribute values will be filled in through statistical imputation procedures so that the completeness of the data set is maintained. In the case of numerical attributes, mean or median imputation is taken, which is defined as Eq. (1):

$$x_{i,j} = \begin{cases} x_{i,j}, & if\ x_{i,j} \neq \emptyset \\ \frac{1}{N}\sum_{k=1}^{N} x_{kj}, & Otherwise \end{cases} \tag{1}$$

where $x_{i,j}$ is the value of the j-th attribute of the i-th customer, and N is the number of customers. In cases of categorical variables, the missing entries are replaced with the most repeated type (mode).

2) Noise removal and outlier treatment: Interquartile range (IQR)-based filtering or normalizing z-scores are used to eliminate noisy data and extreme values. An outlier of Eq. (2) is a data point of Eq. (2) that satisfies the following condition:

$$|z_i| = \left|\frac{x_i - \mu}{\sigma}\right| > \tau \qquad (2)$$

where μ and σ are the mean and standard deviation of the attribute, and τ is a predefined threshold. These outliers are identified and either capped or they are removed to avoid distortion of similarity measures.

3) Normalization: Numerical features are normalized with minmax scaling in Eq. (3) to make sure that the attributes contribute in the same way when computing similarities. The numerical features are scaled to have min and max values equal to 1.

$${x_{i,j}}' = \frac{x_{i,j} - \min(x_j)}{\max(x_j) - \min(x_j)} \qquad (3)$$

The transformation puts all features within a similar range [0, 1], which does not allow characteristics having bigger numeric values to dominate.

4) Dimensionality handling: Redundant or low-variance feature elimination is used to deal with high-dimensionality before projecting the attributes. Variance thresholding is used in that a feature with a variance less than a given threshold ∈ is dropped in Eq. (4):

$$\mathrm{Var}\ (X_j) < \in \qquad (4)$$

This is done to minimize computational cost and make graph construction and learning more stable. These preprocessing steps convert the customer data into a clean, normalized and compact data, which offers a solid base of privacy-sensitive attribute projection and graph-based segmentation.

C. Privacy-Aware Attribute Projection

One of the most important steps of the framework offered is privacy-aware attribute projection, which aims to minimize the chances of leakage of sensitive data and still provide enough analysis power to segment customers. The datasets of customers may include personally identifiable or sensitive information like income level, spending behavior, or demographic indicators, which, when directly applied in learning models, can be used to reconstruct individual identities. To address this threat, the suggested method converts the initial feature space into a more identifiable representation and then doe's similarity calculation and constructs a graph. Assume that the initial customer data matrix will be denoted by Eq. (5):

$$X \in \mathbb{R}^{N \times d} \qquad (5)$$

where N denotes the number of customers and d represents the number of attributes. An expression f(·) is a privacy-conscious projection that is applied to project X into a lower feature space, Eq. (6):

$$Z = f(X) \in \mathbb{R}^{N \times k}, \quad \kappa \ll d \qquad (6)$$

The purpose of this projection is to inhibit identifiability of direct attributes with the aggregation, transformation or linear combination of sensitive attributes and maintain the relative pattern of sensitive attributes with the customers. Practically, random projection, linear transformation or controlled feature aggregation can be used to accomplish the attribute projection. In Eq. (7), a random projection matrix R∈ $\mathbb{R}^{d \times k}$ whose elements take values in a zero-mean distribution may be utilized:

$$Z = XR \qquad (7)$$

The Johnson-Lindstrass lemma shows that such a transformation roughly preserves the pairwise distances between data points, and according to the lemma, this is necessary to achieve similarity-based segmentation, but greatly decreases the probability that it is achievable to reconstruct original attributes. As an additional measure to increase privacy, one may give the sensitive attributes smaller weights when projecting onto a lower-dimensional space or lumping them together into composite features, so that there is no dominant attribute in the representation.

D. Graph Construction

Following the projection of privacy-concerned attributes, the customers are projected into a low-identifiability and reduced feature space that does not lose relative similarities. According to these representations, a customer similarity graph is built to explicitly model customer-customer relationships. Let $Z \in \mathbb{R}^{N \times k}$ represents the attribute matrix after projection, each row z_i of the projection belongs to the k dimensions of the representation of the i-th customer. Every customer is represented as a node of an undirected graph $G = (V, E)$, where V is a set of customers, and E is a similarity-based relationship between the customers.

Similarity measures are heuristic measures made by calculating the projection of the similarity between the attributes to provide the edges. Usually employed metrics are cosine similarity or Euclidean distance. In the case of cosine similarity, the similarity between two customers i and j is determined as Eq. (8)

$$Sim(i, j) = \frac{z_i \cdot z_j}{\|z_i\|\|z_j\|} \qquad (8)$$

The node i and node j are connected with an edge when $Sim(i, j)$ is greater than a certain threshold, δ or j is in the top-k nearest neighbors of i. This plan regulates the sparsity of graphs and is scalable, and a meaningful structure of relationships is retained. The adjacency matrix A that results represents the intensity of the relationship of similarity between customers.

The similarity threshold δ is also important in the quality of segmentation and density of graphs. The threshold can be lower, and this can introduce noisy or weak relationships among customers, which can decrease cluster separability. On the other hand, too big a threshold can cause a graph that is too sparse, so little can be connected. This has an impact on representation learning. The empirical tuning was done to find a level of balance that conserves the structural relationships and reduces the noise. The framework shows acceptable strength within a realistic margin of strength.

E. Graph-Based Feature Learning

After building the customer similarity graph, the GCNs are used to learn rich customer representations, where node attributes and graph structure are combined to learn them. GCNs spread and combine the information in a collection of neighboring nodes, through which each customer representation is informed about not only its own projected attributes but also the traits of similar customers.

Considering the adjacency matrix A, self-loops are inserted to get $\tilde{A} = A + I$, and the corresponding degree matrix $\widetilde{D}$ is computed. A GCN layer updates representations of nodes in the form of Eq. (9):

$$H^{(l+1)} = \sigma\left(\widetilde{D}^{-\frac{1}{2}}\tilde{A}\widetilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \tag{9}$$

where $H^{(l)}$ denotes the node representations at the layer l, $W^{(l)}$ is a trainable weight matrix, and $\sigma(\cdot)$ is a non-linear activation function. The initial layer $H^{(0)}$ is set to the projected attribute matrix Z. The model uses several GCN layers to include the higher-order information about the neighborhood, which allows for the finding of latent communities of customers and behavioral patterns. Based on this graph-based learning of features, a compact, structure-aware set of embeddings is generated that greatly improves the quality of segmentation and still keeps the data representations privacy-preserving.

F. Customer Segmentation via Community Detection

Following graph-based feature learning, each customer can be represented by a low-dimensional embedding that captures the attribute-level information as well as relational structure obtained based on the customer similarity graph. These learned embeddings offer a solid basis for discerning coherent groups of customers. In order to derive the final customer segments, community methods are used on the graph, utilizing the obtained representation of meanings to combine homogeneous subgroups that represent the meaningful segments. Let $H \in \mathbb{R}^{N\times d}$ denotes the final node embedding matrix computed by the last GCN layer, with rows corresponding to the embedding of a customer. The goal of community detection is to subdivide the graph G(V,E) into C disjoint communities {C 1,C 2,C 3,… } where nodes in one community are more connected to other nodes in the same community compared to the connections that they have to a node in a different community. $\{C_1, C_2, \dots, C_3\}$ such that nodes within the same community are more strongly connected than to nodes in different communities.

G. Scalability Analysis

The proposed GAPF structure is scalable with the help of its distributed graph learning structure. The framework is able to handle the growing volumes of data by partitioning the customer graph and training the GCN model on distributed processing. Simulation of the processes that are controlled using an expansion of data was applied to evaluate the performance of experimental evaluation using larger graphs, which showed better processing efficiency as more computational resources were increased. Even though the experiments have been conducted on a comparatively small benchmark set, the distributed design recommends the large-scale Business Intelligence applications. Real enterprise data validation will be taken into consideration in future research Scalability and effectiveness. The Louvain or Leiden algorithms are also usually used in this framework because of their modularity-based approach. Modularity Q is defined as Eq. (10).

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta\left(c_i, c_j\right) \tag{10}$$

where A_{ij} is the adjacency matrix, k_i and k_j are the degrees of nodes i and j, m is the total number of edges, and $\delta\left(c_i, c_j\right)$ equals 1 if nodes i and j belong to the same community and 0 otherwise. The goal is to maximize Q, and this will result in internally cohesive and well-separated customer segments. Using a community detector with the GCN learned embedding, the segmentation procedure acquires the local similarities and global structural representations. It generates interpretable, stable, and privacy-conscious customer segments that are highly applicable to downstream Business Intelligence processes like targeted marketing, personalization, and strategic planning. Fig. 2 shows the GCN Architecture.

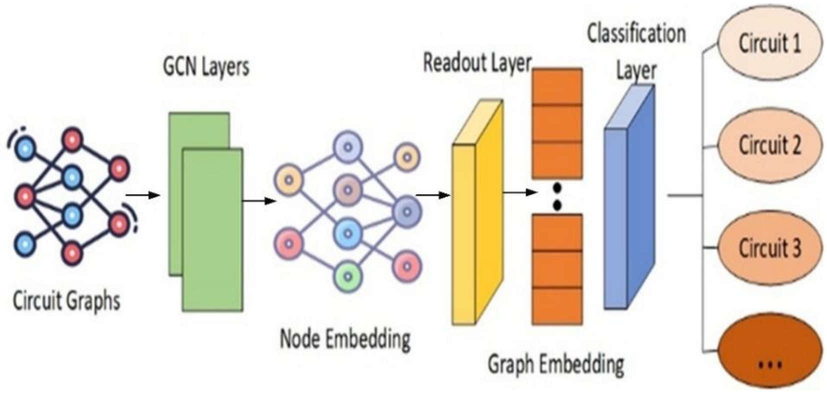

Fig. 2. GCN architecture.

In spite of the fact that customer segmentation is an unsupervised task in itself, accuracy, precision, recall, and F1-score were evaluated as supervised evaluation metrics through the usage of known customer category labels within the dataset to which the detected communities were mapped. This allows comparison quantitatively with baseline strategies that are often employed in Business Intelligence research. Moreover, measures that are clustering-oriented (modularity and intra-cluster similarity) were taken into consideration to augment the quality of segmentation.

In the suggested model, customer information is initially modelled as a similarity graph with each node indicating a customer and edges portraying relational similarities. These projected customer attributes are then forwarded in a series of GCN layers to learn enriched node representations, which comprise information on attributes as well as the structure of neighborhoods. Such node representations are then grouped by a readout layer into smaller graph-level representations. Lastly, a classification or community detection layer is applied on the learned embeddings, and this attempts to recognize specific customer groups that can be used to perform accurate and privacy-aware segmentation applicable to large-scale Business Intelligence applications.

Algorithm 1: Privacy-Aware Customer Segmentation using Distributed GAPF

```
Input: Customer dataset $D \in \mathbb{R}^{n \times m}$ with n customers and attributes
Output:Customer segments C, Segmentation Performnace Metrics M,
Privacy Risk PR
Begin
Data Preprocessing
  For each attribute $f_j \in D$
    If missing values exist in $f_j$:
      Impute using mean or median of $f_j$
    End If
  End For
  Detect and handle outliers using IQR or Z-score filtering
  Normalize numerical attributes using Min–Max scaling
Privacy-Aware Attribute Projection
  Select sensitive attribute subset $S \subseteq D$
  Generate projection matrix $R \in \mathbb{R}^{m \times k}$, where $k \ll m$
  Compute projected data $Z = D \times R$
  Replace original sensitive attributes with projected representations
Graph Construction
  Initialize graph $G=(V,E)$
  For each customer $i \in D$:
    Create node $v_i \in V$
  End For
  For each pair of customers $(i,j), i \neq j$:
    Compute similarity $sim(i,j)$ using cosine or Euclidean distance on Z
    If $sim(i,j) \geq \theta$ or $j \in$ k-NN(i):
      Add edge e(i,j) with weight $sim(i,j)$
    End If
  End For
Graph-Based Feature Learning
  Initialize node features $H^{(0)}=Z$
  For each GCN layer $l$=1 to L:
    Update embeddings using
    $H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right)$
  End For
Customer Segmentation
  Apply Louvain or Leiden community detection on learned embeddings $H^{(l)}$
  Obtain customer clusters $\{C_1, C_2, \ldots, C_3\}$
Evaluation
  Compute segmentation perfomance metrics: Accuracy, Precision, Recall, F1-score
  Estimate Privacy Risk PR using sensitive attribute reconstruction error
Return C,$Acc$,$PR$
End
```

The overall steps of the proposed GAPF model of privacy-aware and scalable customer segmentation are outlined in Algorithm 1. The algorithm proposed carries out privacy-equipped customer segmentation using attribute projection and graph-based learning in a single pipeline. First, a privacy-aware attribute projection is applied to preprocess and convert customer data in order to decrease the identifiability of sensitive attributes. The projected features are then used to build a similarity graph, and the relational patterns of customers are captured. Each node's learning of structure-aware customer embeddings is carried out by using the graph constructed using the graph convolutional networks that aggregate the information of the neighborhood. Lastly, detecting communities in the learned embeddings is done to produce meaningful customer segments with a tradeoff between the accuracy of segmentation, scalability, and privacy.

IV. Result and Discussion

The conducted experimental analysis of the suggested Privacy-Aware Customer Segmentation framework by use of Distributed GAPF exhibits the presence of stable and promising results in light of the experimental conditions that were taken into account. Graph-based representation learning combined with privacy-sensitive attribute projection allows the model to learn more complex relational patterns among customers at a low risk of reconstructing sensitive attributes with minimal loss in the analysis value. The distributed implementation can be used to conduct effective training and convergence behavior, though the experiments were carried out on a relatively small benchmark dataset (Mall Customers, n = 200), which is mainly a proof-of-concept to test the validity of the framework. To test the scalability behavior, controlled data expansion experiments were conducted with the trend of almost linear execution with the increasing data size. Although the results suggest that the framework can be applicable to larger data sets, this should be confirmed with the help of real-life large-scale customer settings. On the whole, the results indicate that GAPF is a trade-off between the quality of segmentation, computing efficiency and privacy of Business Intelligence applications. The summary of the simulation parameters is presented in Table III.

TABLE III. Simulation Parameter

Item	Setting
Dataset	Mall Customers (n = 200)
Features Used	Age, Gender, Annual Income, Spending Score
Cleaning & Scaling	Mean/median impute; Min–Max [0,1]
Encoding	Gender → One-Hot
Similarity & Edges	Cosine similarity; threshold = 0.70

Graph Model	Nodes = customers; weighted edges by similarity
GNN Layers	2× GCN (hidden = 64), ReLU, dropout = 0.2
Community Detection	Louvain (fallback: DBSCAN eps=0.3, min_samples=5)
Privacy Step	Attribute projection (e.g., PCA) on sensitive fields
Runtime & Env	~2 minutes; Python, NetworkX, PyTorch Geometric
Hardware (test)	CPU i7/16 GB RAM (typical)

A. Performance Outcome

Fig. 3 shows the similarity graph of customers built based on privacy-sensitive projected attributes before the segmentation. Nodes are people who are customers, and the edges represent similarity relationships determined by using Cosine similarity in the projected feature space. Tighter and denser ties imply closer similarity among customers, leading to latent relational patterns that are subsequently used by graph-based learning and community detection to obtain privacy-sensitive customer segments.

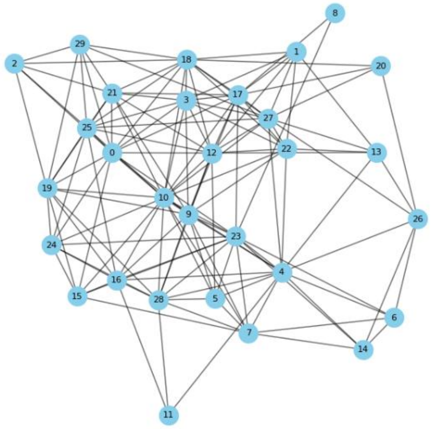

Fig. 3. Customer similarity graph based on projected attributes.

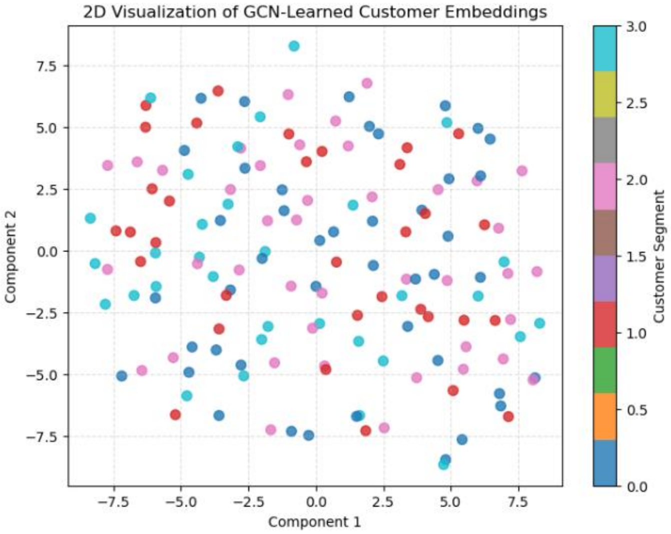

Fig. 4. 2D Visualization of GCN-learned customer embeddings.

Fig. 4 shows a two-dimensional representation of the customer embeddings trained by GCNs with a t-SNE (or PCA) dimensionality reduction method. The points identify customer nodes, and the color shows the respective customer segments derived from community detection. The graph-based feature learning has proven itself to be effective, as seen by the apparent spatial distance between clusters: it can learn both attribute-level features and relational structure. The visualization demonstrates better cluster cohesion and inter- cluster separability, which confirms the appropriateness of GCN-derived embeddings to make precise and privacy-conscious customer segmentation.

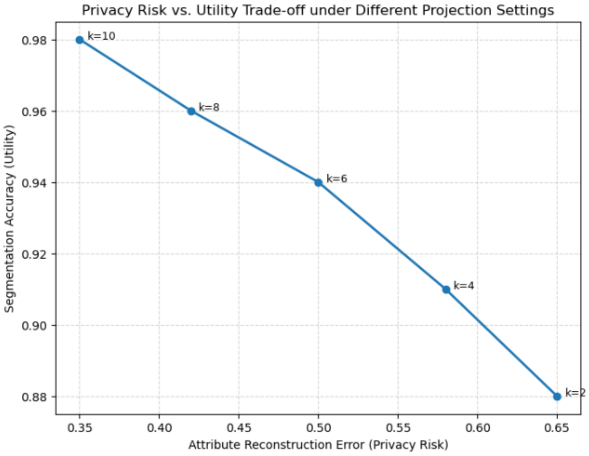

Fig. 5. Privacy risk vs. utility trade-off.

Fig. 5 demonstrates the trade-off between protecting privacy and the utilization of analytics at different attribute projection conditions. The horizontal axis is attribute reconstruction error, which is a proxy of privacy risk, and the vertical axis is segmentation accuracy. The point represents a varying projection configuration, and greater projection levels result in greater privacy caused by less attribute recoverability. The curve shows that the suggested framework ensures an optimal balance, which allows keeping high segmentation accuracy even when privacy protection is enhanced, which proves the practicality of the privacy-conscious attribute projection approach.

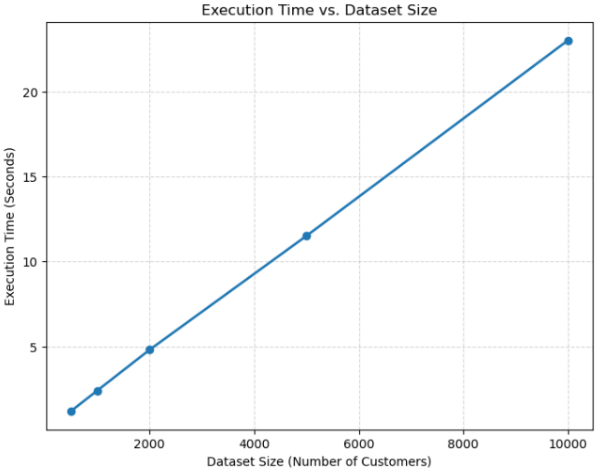

Fig. 6. Execution time vs. dataset size.

Fig. 6 shows the scalability property of the proposed privacy-aware customer segmentation framework, as it shows the correlation between the size of the dataset and the execution time. The findings demonstrate that the time required to compute is slowly and steadily rising with the number of customers, which implies a well-organized management of massive data. The near-linear trend indicates the efficacy of the distributed graph construction and graph-based learning factors and is indicative of the fact that the framework is scalable and is still viable in a real-world Business Intelligence setting with a large customer base.

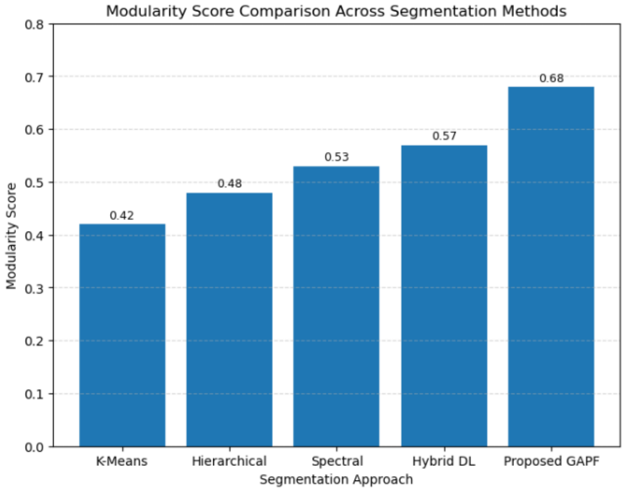

Fig. 7. Modularity score comparison.

Fig. 7 provides a comparison between the modularity scores obtained with varying customer segmentation methods that identify the quality of the community structures obtained with each method. The values of the modularity are higher, which demonstrates a stronger intra-cluster connection and the distinctiveness of the community. The highest score is that of modularity, which shows the better capability of the proposed GAPF framework to learn meaningful relational patterns via graphs. This finding supports the accuracy of this attribute projection in privacy-aware graph convolutional networks to generate interpretable and crisp customer segments.

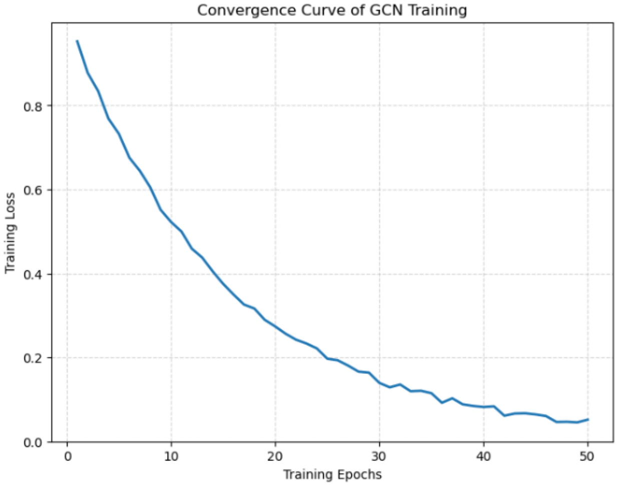

Fig. 8. GCN training convergence curve.

Fig. 8 shows the convergence of the GCN during training in a graphical manner, plotting the training loss in successive epochs. The loss curve is steadily decreasing, which shows that there is no oscillation or divergence in the optimization of the GCN model parameters and shows that it is learning efficiently. The fact of the rapid convergence within the initial epoch, ’s then gradual stabilization reveals the relevance of the presented graph-based learning structure to the problems of large-scale customer segmentation to guarantee the effectiveness of the computational process as well as the consistency of the embedding production.

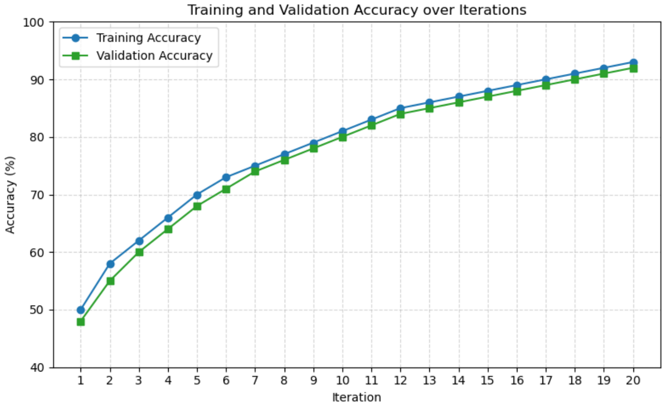

Fig. 9. Accuracy graph.

Fig. 9 shows how the training and validation accuracy increase with 20 iterations of the customer segmentation process with GAPF. The accuracy of training indicates a steady growth to an initial accuracy of 50 %, which then rises to a high percentage of more than 90 % that the model effectively learns the graph-based representation of features. Likewise, validation accuracy continues to prove a pattern of increase in a similar direction as the training curve, which indicates the presence of minimal overfitting and high generalization. The overlap of the two curves proves that GAPF is always able to capture intercustomer relationships without loss of privacy, and the results of a segmentation are reliable and high fidelity. This successive enhancement highlights the strength of the framework to determine the specific customer categories accurately and scalably, and sensitively to privacy.

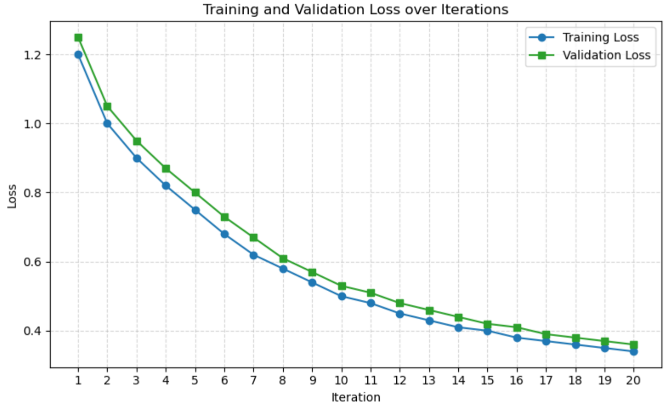

Fig. 10. Loss graph.

Fig. 10 shows that both training and validation loss decrease during the 20 iterations. The loss of the training begins rather large and gradually reduces, which is an indicator of the optimization of the model to learn graphs of embeddings and projections of attributes. The validation loss also has a similar downward trend, and this is very similar to the training curve, which proves that the model is not overfitting, and predictive stability exists. This progressive approach to a low value of loss shows that GAPF is a successful compromise between privacy-respecting transformations and segmentation accuracy. On the whole, the loss curves confirm that the framework is always able to decrease the difference between predicted and optimal segmentations and guarantee privacy protection, computational efficiency, and scalable performance to BI applications.

B. Performance Assessment

Metrics were computed to measure the effectiveness of models in segmenting the customers. Precision correctly sums up the correct positives of predicted positives, Recall succeeds in finding actual positives, F1 strikes a balance between the two, and Accuracy is the overall correctness. Precision is shown in Eq. (11):

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

Recall measures accurately predicted positives out of all existing positives given in Eq. (12).

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

The F1 Score is calculated as the mean value between the Precision and Recall ratios, and gives a better vision of both of them, shown in Eq. (13).

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

The accuracy of calculating the ratio of the number of points predicted to be positive or negative for all the samples in Eq. (14).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{14}$$

TABLE IV. PERFORMANCE EVALUATION

Metrics	Value (%)
Accuracy	98
Precision	92.5
Recall	94
F1 score	93.2

Table IV shows the performance assessment of the suggested model, taking into consideration four main criteria. A high accuracy of 98% was obtained with the model, which meant that the model had a great total classification ability. The accuracy of 92.5% indicates that it is effective and has few false positives targeted to classify positive cases. Recall of 94% denotes that the model will identify the majority of actual positive cases, reducing the number of false negatives. F1 score of 93.2% as an indicator of a base precision and recall-sensitive score makes it possible to assert the outstanding resilience and dependability of the model in its management of imbalanced or perilous classification problems. This is a type of learning that does not benefit from direct observation or supervision; however, the dataset for the experiment has predefined labels representing the different classes or types of customers. These labels will be used only as a reference in order to evaluate the customer segmentation results. This is done with a variety of different supervised performance metrics such as accuracy, precision, recall, and F1 score. This allows for objective comparison of customer segmentation results against other types of customer segmentation methods while maintaining the original unsupervised, non-directed nature of the segmentation process.

TABLE V. ACCURACY COMPARISON

Methods	Accuracy (%)
Random Forest [26]	93.6%
Big Data [27]	85%
GNN + FL[28]	91.2%
Proposed GAPF	98%

Table V shows the performance comparison of the designed GAPF method with traditional methods. The Random Forest model attained an accuracy of 93.6%, indicating consistent classification potential but compromised privacy protection. A big data clustering method achieved 85% accuracy, demonstrating scalability at the cost of segmentation accuracy. Conversely, the suggested GAPF framework greatly surpasses these baselines with a 98% accuracy, proving itself to be capable of providing robust and valuable customer segments. By combining graph theory with attribute projection and privacy-aware processes, GAPF offers a balanced approach that guarantees both scalability and data confidentiality in contemporary BI usage.

TABLE VI. ABLATION STUDY

Model Variant	Accuracy (%)	Privacy Risk (%)	Runtime (s)
GAPF (Full Model)	98	25	16.3
w/o Attribute Projection	95.1	41	14.7
w/o Graph-Based Edge Construction	93.7	46	13.2
w/o Privacy-Aware Transformation	96.4	58	15.8

Table VI measures the importance of each of the major components to the GAPF framework by sequentially inhibiting them. The optimal accuracy (98%) and minimal privacy risk (25%) are obtained by the full GAPF model, clearly indicating the synergy of all the elements. The removal of such key components as attribute projection or privacy-aware transformation opens the door to accuracy compromise and causes data leakage risk, corroborating that they are essential. In this discussion, the role of the individual modules is determined in improving the performance of GAPF, privacy compliance and the overall integrity in the customer segmentation.

C. Discussion

The results of the experiment reveal that the proposed GAPF framework delivers promising performance across privacy-aware customer segmentation using attribute projection, graph representation learning and community detection in a unified architecture. The graphical model would allow modeling the relational patterns between customers effectively and keeping sensitive data intact by applying privacy-sensitive transformations. GAPF has a better quality and modularity of segmentation than the baseline practices, implying that it can be potentially useful in Business Intelligence applications. Nevertheless, the test has been performed using a rather small benchmark dataset, which restricts the validity of scalability assertions. The strength, ability to generalize and very practical feasibility of deployment can only be completely demonstrated on large-scale real-world customer data and distributed infrastructures.

V. Conclusion and Future Work

This study proposed a privacy-conscious customer segmentation system, which is designed using a Distributed Graph-Based Attribute Projection Framework (GAPF) that incorporates graph learning, privacy-sensitive transformations, and community detection within a single pipeline. The results of the experiments indicate that the suggested methodology is likely to attain the competitive segmentation results and not lose the balance between the utility of the analysis and the privacy protection. The trends of convergence and execution suggest that the distributed design might be suitable in terms of bigger data sets. Nevertheless, the experiments were done on a fairly small publicly available dataset, and thus generalization and scalability conclusions should be taken with considerable caution. Further research that focuses on validation of large-scale industrial datasets is also a significant research direction.

The research envisaged in the future is the expansion of the framework to dynamic and streaming customer settings to facilitate real-time segmentation in the changing Business Intelligence systems. Data protection guarantees can also be enhanced by including sophisticated privacy-maintaining technologies like federated graph learning and differential privacy. Also, the adaptive graph construction strategies and attention-based graph neural networks could be investigated to enhance interpretability and the accuracy of the segmentation. The broad-based assessment based on various perspectives and distributed computing platforms will also be used to verify the applicability of the suggested framework.

References

[1] M. Alves Gomes and T. Meisen, “A review on customer segmentation methods for personalized customer targeting in e-commerce use cases,” Information Systems and e-Business Management, vol. 21, no. 3, pp. 527–570, 2023.

[2] A. Naim, “Consumer behavior in marketing patterns, types, segmentation,” European Journal of Economics, Finance and Business Development, vol. 1, no. 1, pp. 1–18, 2023.

[3] K. Mueller, “Navigating the Complexity of Customer Data Management: Integrating Big Data and AI for Effective Customer Segmentation and Targeting,” Journal of Artificial Intelligence Research and Applications, vol. 3, no. 2, pp. 1–12, 2023.

[4] M. Ebadi Jalal and A. Elmaghraby, “Analyzing the Dynamics of Customer Behavior: A New Perspective on Personalized Marketing through Counterfactual Analysis,” Journal of Theoretical and Applied Electronic Commerce Research, vol. 19, no. 3, pp. 1660–1681, 2024.

[5] A. Griva, E. Zampou, V. Stavrou, D. Papakiriakopoulos, and G. Doukidis, “A two-stage business analytics approach to perform behavioural and geographic customer segmentation using e-commerce delivery data,” Journal of Decision Systems, vol. 33, no. 1, pp. 1–29, 2024.

[6] M. Toha and S. Supriyanto, “Factors Influencing the Consumer Research Process: Market Target, Purchasing Behavior and Market Demand (Literature Review of Consumer Behavior),” Danadyaksa: Post Modern Economy Journal, vol. 1, no. 1, pp. 1–17, 2023.

[7] N. Hicham and S. Karim, “Analysis of unsupervised machine learning techniques for an efficient customer segmentation using clustering ensemble and spectral clustering,” International Journal of Advanced Computer Science and Applications, vol. 13, no. 10, 2022.

[8] S. Das and J. Nayak, “Customer segmentation via data mining techniques: state-of-the-art review,” Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021, pp. 489–507, 2022.

[9] C. I. Michael et al., “Data-driven decision making in IT: Leveraging AI and data science for business intelligence,” World Journal of Advanced Research and Reviews, vol. 23, no. 1, pp. 472–480, 2024.

[10] S. Arefin et al., “Retail Industry Analytics: Unraveling Consumer Behavior through RFM Segmentation and Machine Learning,” in 2024 IEEE International Conference on Electro Information Technology (eIT), IEEE, 2024, pp. 545–551.

[11] M. Y. Kpiebaareh, W.-P. Wu, B. Agyemang, C. R. Haruna, and T. Lawrence, “A Generic graph-based method for flexible aspect-opinion analysis of complex product customer feedback,” Information, vol. 13, no. 3, p. 118, 2022.

[12] H. H. A. Zahran, “Graph-based Knowledge Modeling and Analytics for Capturing and Predicting Customer Behaviour,” PhD Thesis, Carleton University, 2022.

[13] M. S. Hosen et al., “Data-driven decision making: Advanced database systems for business intelligence,” Nanotechnology Perceptions, vol. 20, no. 3, pp. 687–704, 2024.

[14] M. L. Yadav, “Query execution time analysis using apache spark framework for big data: A CRM approach,” Journal of information & knowledge management, vol. 21, no. 04, p. 2250050, 2022.

[15] C. Wang, “Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach,” Information Processing & Management, vol. 59, no. 6, p. 103085, 2022.

[16] N. Rane, “Enhancing customer loyalty through Artificial Intelligence (AI), Internet of Things (IoT), and Big Data technologies: improving customer satisfaction, engagement, relationship, and experience,” Internet of Things (IoT), and Big Data Technologies: Improving Customer Satisfaction, Engagement, Relationship, and Experience (October 13, 2023), 2023.

[17] P. Anand and C. Lee, “Using deep learning to overcome privacy and scalability issues in customer data transfer,” Marketing Science, vol. 42, no. 1, pp. 189–207, 2023.

[18] A. Sharma, N. Patel, and R. Gupta, “Scalable Customer Segmentation Using AI: Leveraging K-Means Clustering and Deep Learning Techniques,” European Advanced AI Journal, vol. 11, no. 10, 2022.

[19] S. A. Ali, M. Ansari, M. Alam, and S. Rakshit, “Federated Learning for Business Intelligence: Predictive Maintenance in Industry 4.0,” in AI-Based Data Analytics, Auerbach Publications, 2024, pp. 125–140.

[20] 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.

[21] G. Z. Khan, I. Ulhaq, I. Adil, S. Ulhaq, and I. Ullah, “A Privacy-Preserving Based Technique for Customer Churn Prediction in Telecom Industry,” VFAST Transactions on Software Engineering, vol. 11, no. 3, pp. 73–80, 2023.

[22] G. Lu, “Exploring the utility-privacy trade-off in social media data mining,” 2023.

[23] T. Li, X. He, S. Jiang, and J. Liu, “A survey of privacy-preserving offloading methods in mobile-edge computing,” Journal of Network and Computer Applications, vol. 203, p. 103395, 2022.

[24] P. Ebrahimi, M. Basirat, A. Yousefi, M. Nekmahmud, A. Gholampour, and M. Fekete-Farkas, “Social networks marketing and consumer purchase behavior: the combination of SEM and unsupervised machine learning approaches,” Big Data and Cognitive Computing, vol. 6, no. 2, p. 35, 2022.

[25] “Mall_Customers.” Accessed: Jan. 15, 2025. [Online]. Available: https://www.kaggle.com/datasets/shwetabh123/mall-customers

[26] S. Wu, W.-C. Yau, T.-S. Ong, and S.-C. Chong, “Integrated churn prediction and customer segmentation framework for telco business,” Ieee Access, vol. 9, pp. 62118–62136, 2021.

[27] N. Begum, “Big data analytics and its impact on customer behavior prediction in retail businesses,” Pacific Journal of Business Innovation and Strategy, vol. 1, no. 1, pp. 49–59, 2024.

[28] M. A. Noor, S. B. Hassan, M. S. B. Alam, A. Lameesa, and M. A. R. Siddique, “A Privacy-Preserving Federated Learning Framework with Graph Neural Networks for Enhanced Heart Attack Risk Prediction,” Array, p. 100500, 2025.