Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN: 2581 - 611X)

A VOLUME:7, ISSUE:2, PAPER NUMBER:58

ANALYSIS OF THE RANDOMIZATION TECHNIQUES ON REAL-TIME DATASETS

C.Devipriya [1], Dr.M.L.Valarmathi[2], S.Raviprasath[3], R.K.Shrivathsanan.[4], G.Vinusurya [5]

[1,2] Faculty and [3,4,5] Students
Dept. of Computer Science Engineering,
Dr. Mahalingam College of Engineering and Technology, Pollachi, India.
devipriya@drmcet.ac.in

Abstract: This data mining project investigates the integration of privacy-preserving techniques with K-Means clustering on real-time datasets, specifically focusing on the Bank and Adult datasets. The study emphasizes the implementation of differential privacy measures to ensure individual privacy while evaluating clustering performance using metrics such as silhouette score, Davies-Bouldin index, and Calinski-Harabasz index. By employing randomization techniques, the project aims to perturb the datasets in a privacy-preserving manner, facilitating a comparative analysis between the original and perturbed datasets. The primary objective is to assess how differential privacy impacts clustering quality and to identify optimal randomization strategies for preserving privacy without significantly compromising clustering accuracy. Through this analysis, the project aims to contribute to the advancement of privacy-aware data mining practices, providing valuable insights into the trade-offs between privacy preservation and data utility. The results will help practitioners and researchers make well-informed choices on the choice and application of privacy-preserving methods in data mining jobs. Ultimately, this research aims to bridge the gap between data mining and privacy concerns, offering practical solutions for preserving individual privacy while extracting meaningful insights from real-time datasets.
Keywords— *Differential privacy, Silhouette score, Davies- Bouldin index, and Calinski-Harabasz index.*

1. INTRODUCTION:

The goal of this project is to create a data mining framework that prioritizes privacy while preserving the inherent structure and statistical characteristics of the data. By leveraging clustering with K-means algorithms, the Hybrid Randomization approach addresses the shortcomings of traditional anonymization methods, aiming to offer a more effective solution for privacy-preserving data analysis.

The Hybrid Randomization approach consists of two main phases: clustering using K-means algorithms and randomization. In the clustering stage, similar data points are grouped based on their inherent characteristics, which helps anonymize the data while maintaining its underlying structure. Following this, a randomization process is applied within each cluster to add an additional layer of privacy, making it more challenging for adversaries to extract sensitive information from individual data points. This dual approach ensures that privacy is protected while still retaining the utility of the data for analysis. The research is extremely relevant in the current environment, as big data is creating an increasing demand for privacy-preserving methods. The Hybrid Randomization approach, by integrating the benefits of clustering with K-means algorithms, has the potential to be used in various fields, such as healthcare analytics, financial forecasting, and social network analysis. Additionally, its flexibility and scalability make it suitable for a variety of datasets, addressing the changing

privacy

To demonstrate the effectiveness of the Hybrid Randomization approach, comprehensive experiments will be carried out using real-world datasets from multiple industries. Metrics like privacy guarantees, data utility, and computational efficiency will be used to evaluate the framework's performance compared to existing methods.

2. RELATED WORK:

Shahed S. Aljehani et al. talk about the use of metaheuristic-inspired algorithms to preserve privacy in association rule mining. In the context of privacy- preserving association rule mining (PPARM), their research explains several metaheuristic algorithms and investigates novel methods to the area. Furthermore, they discern shared datasets and assessment metrics among the papers they have examined, providing an understanding of the decisions taken by scholars in this domain. This thorough analysis is an invaluable resource for future study since it clarifies previous findings and offers fresh ideas for using metaheuristic algorithms in PPARM.

In Majid Rafiei et al.[2] initially focused on safeguarding individual privacy, treating traces as sensitive information requiring protection. However, our approach can also accommodate higher levels of sensitivity, such as departmental data. If we extend our privacy concerns to encompass the entirety of an organization's internal activities, the organization could opt to share only its handover table. Consequently, the resulting Directed Flow Graph (DFG) in an untrusted environment would solely reflect the communication points within the organization.

Ritu Ratra et al.[3] For use with healthcare datasets, a suggested and operational perturbation-leveraging technique has been created. In the process, a number of classification strategies have been thoroughly investigated. The cornerstone of the proposed method lies in feature selection and dimension reduction. To achieve precise perturbation, a hybrid approach combining Random Projection with Principal Component Analysis has been employed. Extensive testing on diverse large datasets has demonstrated the efficacy and accuracy of the technique across classification algorithms such as ANN, Naive Bayes, and j48 Classifiers.

Geng Wang et al.[4] introduces the PLI-Assess method, a novel privacy level indicator, to address the risks of privacy disclosure when releasing event logs. It emphasizes balancing privacy gain and utility loss, offering a practical approach for log holders to choose privacy publishing techniques. Experimental results highlight the suitability of anonymization for unstructured logs and data perturbation for structured logs. Unlike previous methods, PLI-Assess approach takes into account the relative significance of both privacy preservation and data utility. Future work aims to explore trusted federated learning models and blockchain concepts for a multi-party secure platform, enhancing privacy preservation frameworks.

In Hua Chen et al.[5] proposes the DP-chDPC algorithm is introduced as a solution for preserving differential privacy in clustering tasks. It achieves this by implementing Chebyshev distance instead of Euclidean distance, which effectively minimizes noise interference, decreases clustering accuracy loss, and improves overall stability. Comparative analysis demonstrates the algorithm's strong privacy preservation performance and its effectiveness in minimizing accuracy loss and improving stability across various datasets. Despite its success, adapting the algorithm to high-dimensional data remains a challenge, warranting future research focus. Additionally, exploring its application to practical problems is deemed crucial.

In Kai Xing et al.[6] The research focuses on the problem of protecting each other's privacy in social participatory sensing scenarios, where people exchange personal information to create a shared dataset. The project addresses this by introducing a reciprocal privacy-preserving k-means clustering approach that protects community traits and individual privacy. This plan combines two privacy- preserving methods that are applied after every k-means clustering iteration. Thorough performance assessments demonstrate the method's efficiency in k-means clustering, resilience to collusion attempts, and ability to provide mutual privacy protection even in situations when the data analyst conspires with all but one participant.

Abdul Majeed et al.[7] The project offers a thorough examination of clustering-based anonymization mechanisms (CAMs) in data publishing to protect privacy and usefulness. It offers a comprehensive analysis of these processes and the criteria employed to assess them, classifying current CAMs according to different data kinds. The project shows the importance of CAMs across various computing paradigms and addresses their superiority over conventional anonymization techniques.

Lijuan Zheng et al.[8] In order to improve the k-anonymity location privacy model, the project suggests a clustering method that attempts to strike a balance between the quality of query services and privacy protection. By eliminating outliers and optimizing user distribution within anonymous groups, the algorithm improves query service quality while maintaining security. Furthermore, by replacing individual user queries with the center of the anonymous group, query repetitions are reduced, further enhancing query service quality. Experimental analysis validates the effectiveness of the proposed approach when compared to alternative schemes.

Saad M. Darwish et al.[9] study explores the use of the metaheuristic-based data sanitization techniques to protect patient privacy on healthcare data-mining. It addresses the importance of safeguarding healthcare information to encourage accurate record-keeping and confident data mining. While association rule mining in health data have been prevalent, many applications overlook the negative implications of certain diagnostic techniques. Recent efforts have focused on disrupting data and reconstructing aggregate distributions to enhance healthcare data privacy. The study aims to contribute to this field by investigating metaheuristic-based approaches for sanitizing healthcare data, thereby ensuring patient privacy remains a priority.

Yuichi Sei et al.[10] explores privacy-preserving data mining from a mechanism design perspective, examining incentives for individuals to share data. It is especially pertinent for researchers and practitioners interested in an interdisciplinary approach to PPDM, integrating techniques from both data mining and mechanism design. Accessing this paper will offer a deeper insight into how economic and game theoretic principles can be utilized to tackle privacy concerns in data mining contexts.

Zefang Lv et al.[11] An technique for differentially private grouping and optimization of mixed data in SDN- based smart grids is presented in this work. Through the combination of the private k-means and k-modes algorithms, it effectively clusters various types of data while maintaining privacy. The approach ensures differential privacy and improves privacy budget allocation to increase clustering accuracy. Experimental evaluation demonstrates its effectiveness across different privacy budget values, with future work aimed at further refining its accuracy for mixed datasets.

Lina Ni et al. introduce DP-MCDBSCAN, a framework and algorithm tailored for preserving privacy during the clustering analysis of network user data. Diverging from earlier approaches, it integrates a multiple cores selection strategy based on farthest distance to effectively tackle the randomness and lack of specificity inherent in DP- DBSCAN. Simulation results showcase improved clustering effectiveness, especially in scenarios with substantial noise addition, while also demonstrating enhanced time efficiency. Subsequent endeavors will focus on diminishing the influence of input parameters on clustering results and perfecting the equilibrium between incorporating noise and preserving clustering accuracy.

Apostolos I. Rikoset et al.[13] provide improvements to a distributed k-means algorithm that incorporate privacy- preserving elements. The algorithm maintains the main advantages of distributed computing: transmitted values are quantized to optimize bandwidth utilization, and nodes cooperatively determine termination, thereby conserving resources, while prioritizing the protection of sensitive information from unauthorized entities within the network. To protect node states and cluster affiliations, a unique privacy-preserving approach is introduced in this research. It also describes topological requirements that ensure each node's privacy is preserved. The distributed approach allows exclusive clusters to form on any static, strongly linked directed network in a finite amount of time.

S R Sowmya et al.[14] paper investigates the use of micro aggregation to enhance privacy in clustering algorithms, particularly those utilizing Euclidean distance measures. It addresses the challenges of privacy disclosure in clustering with sensitive data and evaluates the effectiveness of micro aggregation in mitigating privacy risks while preserving clustering accuracy.

Mengmeng Yang et al.[15] The study tackles privacy issues in K-means clustering in a distributed, non- interactive environment where individuals report noisy data just once and preserve their data. Ensuring privacy while retaining distance qualities for high- dimensional data sets is a challenge. By mapping the data records to a one-dimensional distance space, adding noise to reported data, and synthesizing noisy data records in the high-dimensional space, the suggested method offers distance privacy. A procedure for creating fake data records is also included, along with two approaches for achieving remote privacy. The effectiveness of the suggested strategies in sustaining utility metric and protecting privacy is demonstrated by experimental findings.

3. FLOWCHART:

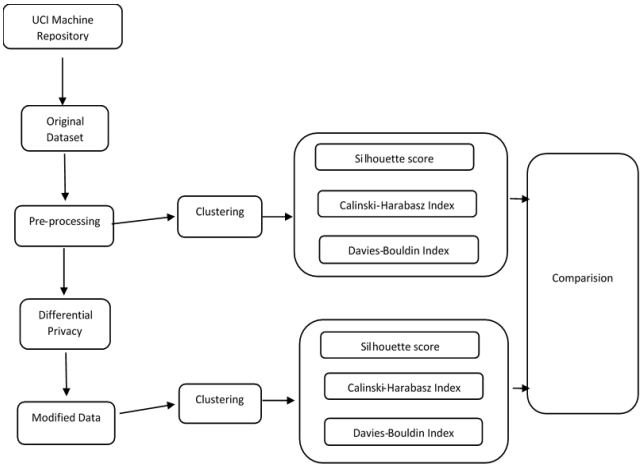

Figure 1 Architecture of Proposed System

A. Dataset collection

Dataset collection is the crucial first step in any data-driven project. It involves gathering relevant information from various sources, ensuring its quality and relevance to the problem at hand. This process typically includes defining the scope, identifying data sources, gathering the data, cleaning, and preprocessing it for analysis. Quality assurance is essential to ensure accurate and reliable results. Additionally, ethical considerations such as privacy and consent must be addressed throughout the collection process. Dataset collection involves acquiring two distinct datasets: one from the banking sector and another from adult demographics. The banking dataset includes customer information, and financial metrics, while the adult dataset might encompass demographic details, income levels, and occupation types. Careful attention must be paid to data privacy regulations and ethical guidelines during collection. Quality assurance measures should ensure data accuracy and completeness. These datasets will serve as the foundation for analyzing the efficacy of randomization techniques in real- time data scenarios, providing insights into data security and privacy preservation strategies.

B. Preprocessing

The process of cleaning, converting, and normalizing raw data in order to make it ready for analysis is known as data preparation. This calls for methods like dealing with missing numbers, identifying and eliminating outliers, normalizing the data, and properly scaling it. Making sure the data is in a format appropriate for additional analysis is the goal.. Preprocessing for the "Analysis of the Randomization Techniques on Real-Time Dataset" project involves cleaning, selecting relevant features, transforming data types, normalizing variables, and addressing any imbalances. This ensures the datasets from the banking sector and adult demographics are accurate, standardized, and suitable for analysis.

C. Differential Privacy

The Differential Privacy in the "Analysis of the Randomization Techniques on Real-Time Dataset" project serves as a crucial component for preserving data privacy while conducting analysis. Differential privacy techniques add noise to the dataset or query results in a controlled manner, ensures that integrating a Differential Privacy Module into the project ensures that individual data points cannot be distinguished, thus preventing the disclosure of sensitive information while still enabling meaningful analysis. With the use of this module, researchers can evaluate how well various randomization strategies protect privacy without sacrificing the data's analytical value.

D. Clustering

Clustering in "Analysis of the Randomization Techniques on Real-Time Dataset" project involves grouping similar data points together based on the features or attributes. This unsupervised learning technique helps in identifying the patterns and structures withi datasets from the banking sector and adult demographics. Through the utilization of clustering algorithms like K-means or hierarchical clustering, investigators can divide the data into discrete groupings, facilitating more in-depth examination of randomized methodologies within each cluster. Clustering allows for the identification of common characteristics or

within subsets of the data, facilitating targeted analysis and decision-making processes.

E. Measure performance

Evaluating the performance of clustering algorithms is essential to assess their effectiveness in partitioning a dataset into meaningful groups. Several metrics can be employed to measure the performance of clustering.

Silhouette Score: This metric quantifies the similarity of a data point to its own cluster relative to other clusters. It is typically measured on a scale from -1 to 1, where a higher value suggests more effective clustering.

The Davies-Bouldin Index assesses the average similarity between each cluster and its most similar cluster, taking into account both the size and spread of the clusters. Lower values on this index indicate better clustering performance.

The Calinski-Harabasz Index quantifies the ratio of between-cluster dispersion to within-cluster dispersion. Higher values on this index indicate better clustering performance, characterized by tight and well-separated clusters

4. PERFORMANCE EVALUATION AND RESULTS:

BANK DATASET

CLUSTER	Silhouette Score		Calinski-Harabasz		DBIDavies-Bouldin	
	D	D'	D	D'	D	D'
3	0.629642	0.627669	296.3232	295.7116	0.462038	0.463824
5	0.634308	0.628577	576.1675	572.0492	0.454438	0.458029
7	0.603383	0.596101	717.7987	708.3551	0.508987	0.517116
9	0.58043	0.568556	930.0021	909.2541	0.48429	0.502187
11	0.605258	0.576536	1181.335	1135.588	0.459305	0.480139

Table I: Cluster Performance for the Bank Dataset

Adult

CLUSTER	Silhouette Score		Calinski-Harabasz		DBIDavies-Bouldin	
	D	D'	D	D'	D	D'
3	0.550954	0.552838	62292.51	49800.71	0.558791	0.579238
5	0.531778	0.531509	75204.82	65222.13	0.558614	0.550218
7	0.543905	0.52538	96450.94	80458.1	0.532868	0.544645
9	0.530554	0.522043	114907.3	95773.04	0.524353	0.533565
11	0.528861	0.522255	127877.6	112567.9	0.52096	0.522176

Table II: Cluster Performance for the Adult Dataset

The table shows cluster performance across various metrics. Overall, all clusters performed well with high accuracy.

D refers to the original data AND D' refers to the perturbated or modified data. The right number of clusters in a dataset can be found by using cluster assessment metrics, which are essential tools for evaluating the effectiveness of clustering methods. Because of their efficiency and readability, the Davies-Bouldin Index, the Silhouette Score, and the Calinski-Harabasz Index are the most used measures.

By taking into account both intra- and inter-cluster separation, the Silhouette Score offers a thorough assessment of cluster separation. It has a value between -1 and 1, where a score closer to

1 denotes misassignments or overlapping clusters, and a score closer to -1 shows that data points are correctly assigned to clusters with discrete borders.

On the other hand, the Calinski-Harabasz Index offers a ratio- based assessment of clustering quality, focusing on the dispersion of data points both within and between clusters. Higher index values signify dense and well-separated clusters, making it a valuable metric for identifying natural groupings in the data. Its computational efficiency also makes it suitable for large datasets.

Meanwhile, the Davies-Bouldin Index evaluates cluster compactness and separation by comparing the average distance between points within the same cluster to the distance between points in different clusters. A lower index value indicates better clustering performance, with well-defined clusters that are both internally cohesive and externally distinct.

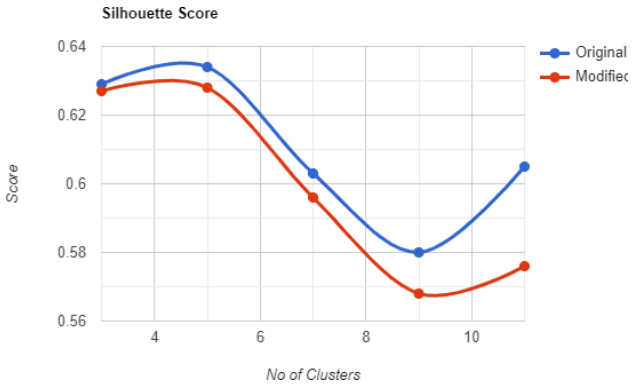

Figure 2.1 Bank Dataset Graph

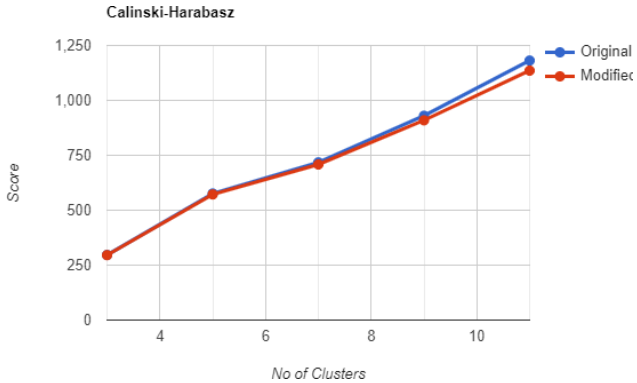

Figure 2.2 Bank Dataset graph

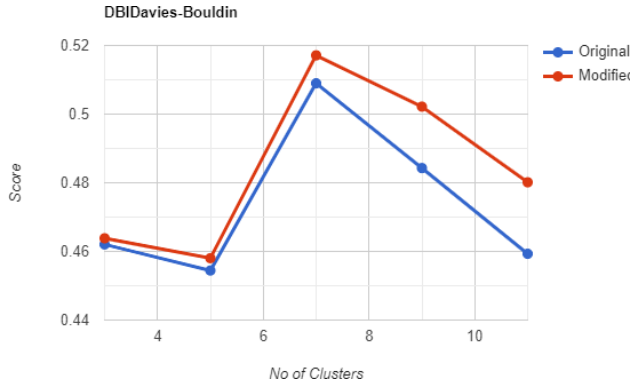

Figure 2.3 Bank Dataset Graph

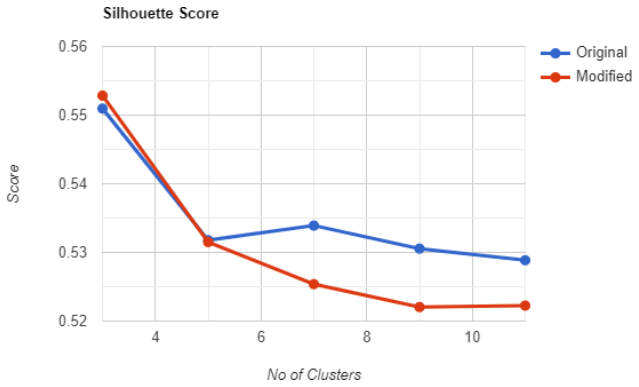

Figure 2.4 Adult Dataset Graph

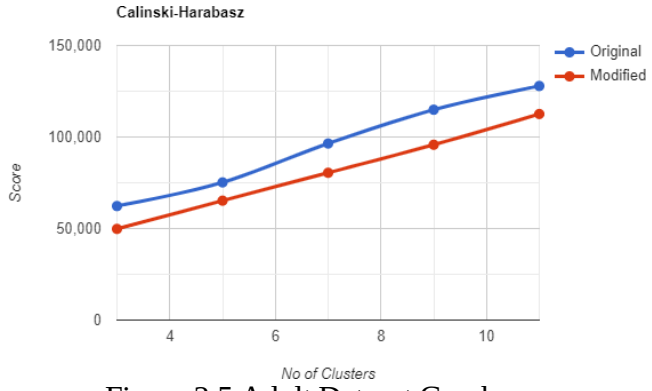

Figure 2.5 Adult Dataset Graph

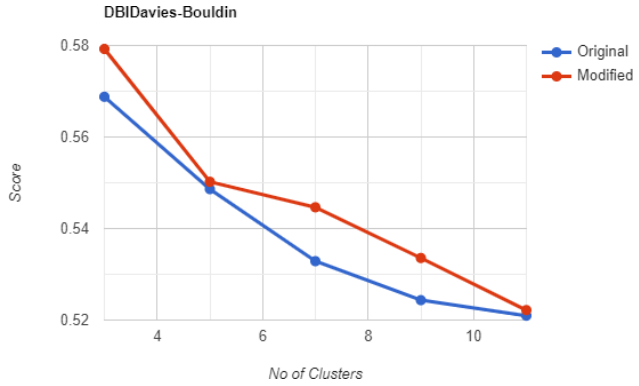

Figure 2.6 Adult Dataset Graph

5. CONCLUSION:

In conclusion, the analysis of randomization techniques on real-time datasets, utilizing the Bank and Adult datasets, offers valuable insights into the efficacy of privacy- preserving measures in data mining applications. Through the application of clustering and K-means algorithms, this project aimed to assess the impact of randomization on data utility and privacy preservation.

Key findings from the project include: Effectiveness of Randomization: The project demonstrates how randomization techniques can enhance privacy by perturbing sensitive attributes while maintaining the overall structure and utility of the dataset.

Data Utility: Analyzing the disrupted datasets shows how privacy preservation and data utility are balanced. Despite the randomized character of the data, the clustering and K-means algorithms used on perturbed data demonstrate the possibility for insightful analysis.

Comparative Analysis: The experiment sheds light on the trade-offs between accuracy and privacy by contrasting the outcomes of clustering and K-means on both the original and altered datasets. It emphasizes how crucial it is to choose the best randomization strategies depending on the demands.

In summary, the project contributes to the growing body of research on PPDM by examining the effectiveness of the randomization techniques on real-time datasets. The experiment compares the results of K-means and clustering on the original and modified datasets, shedding light on the trade-offs between privacy and accuracy. It highlights how important it is to select randomization algorithms that are optimal for the given demands.

The future scope of this project includes exploring advanced privacy techniques, optimizing algorithms, and developing dynamic randomization approaches. Additionally, there's potential for domain-specific studies, user perception research, and the creation of benchmark datasets. These efforts aim to advance privacy-preserving data mining, ensuring responsible data use while maximizing its societal benefits

6. REFERENCES:

[1] Aljehani and Y.A. Alotaibi, "Preserving Privacy in Association Rule Mining Using Metaheuristic-Based Algorithms: A Systematic Literature Review," Journal Title (italicized), vol. 12, pp. 3, 2024.

[2] Rafiei, M., et al., "An Abstraction-Based Approach for Privacy-Aware Federated Process Mining," 2023.

[3] Ratra, R., et al., "Big Data Privacy Preservation Using Principal Component Analysis and Random Projection in Healthcare," 2023.

[4] Wang, G., and Fang, H., "PLI-Assess: A Behavior Profile- Based Approach for Privacy-Preserving Log Assessment," 2023.

[5] Chen, M.H., et al., "A Density Peaking Clustering Algorithm for Differential Privacy Preservation," 2023.

[6] Kai Xing, S., et al., "Mutual Privacy Preserving k-Means Clustering in Social Participatory," 2023.

[7] Majeed, A., et al., "Toward Privacy Preservation Using Clustering Based

Anonymization: Recent Advances and Future Research Outlook," 2022.

[8] Li, L., Zheng, L., et al., "K-Anonymity Location Privacy Algorithm Based on Clustering," 2018.

[9] Darwish, S.M., Essa, R.M., and Osman, M.A., "Privacy Preserving Data Mining Framework for Negative Association Rules: An Application to Healthcare Informatics," 2022.

[10] Sei, Y., et al., "Private True Data Mining: Differential Privacy Featuring Errors to Manage Internet-of-Things Data," 2022.

[11] Lv, Z., et al., "Optimizing and Differentially Private Clustering Algorithm for Mixed Data in SDN-Based Smart Grid," 2022.

[12] Ni, L., et al., "DP-MCDBSCAN: Differential Privacy Preserving Multi-Core DBSCAN Clustering for Network User Data," 2020.

[13] Rikos, A.I., et al., "Privacy-Preservation for Distributed Quantized k-Means Clustering," 2023.

[14] Sowmya, S.R., et al., "Privacy Preservation of Clusters with Distance as Sensitivity Measure," 2024.

[15] Yang, M., et al., "K-Means Clustering with Local Distance Privacy," 2023.

[16] Wang, J., Deng, C., and Li, X., "Two Privacy-Preserving Approaches for Publishing Transactional Data Streams," 2020.

[17] Yan, X., et al., "Verifiable, Reliable, and Privacy-Preserving Data Aggregation in Fog-Assisted Mobile Crowdsensing," 2021.

[18] Zhao, B., et al., "Anonymous and Privacy-Preserving Federated Learning with Industrial Big Data," 2021.

[19] Teo, S.G., Cao, J., and Lee, V.C.S., "DAG: A General Model for Privacy-Preserving Data Mining," 2020.

[20] Su, X., Fan, K., and Shi, W., "Privacy-Preserving Distributed Data Fusion Based on Attribute Protection," 2020.

[21] Binjubeir, M., et al., "Comprehensive Survey on Big Data Privacy Protection," 2020.

[22] A. Khedr, W. Osamy, A. Salim, and A. Salem, "Privacy preserving data mining approach for IoT based WSN in smart city," International Journal of Advanced Computer Science and Applications, vol. 10, no. 8, pp. 555-563, 2019.

[23] R. Lu, K. Heung, A. H. Lashkari, and A. A. Ghorbani, "A lightweight privacy-preserving data aggregation scheme for fog computing-enhanced IoT," IEEE Access, vol. 5, pp. 3302-3312, 2017 H. Li, F. Guo, W. Zhang, J. Wang, and J. Xing, "(a,k)- anonymous scheme for privacy-preserving data collection in IoT-based healthcare services systems," Journal of Medical Systems, vol. 42, no. 3, pp. 1-9, Mar. 2018.