Comparative Analysis of Algorithms for Sensitive Outlier Protection in Privacy Preserving Data Mining

Muhammad Ikhwan Burhan[1,*], Andi Nurfadillah Ali[2], A. Inayah Auliyah[3], Muhaimin Hading[4]

[1,2,3,4] Information System Study Program, Bacharuddin Jusuf Habibie Institute of Technology, Parepare, Indonesia
Email: [1,*]ikhwan@ith.ac.id, [2]anurfadillah@ith.ac.id, [3]inayah@ith.ac.id, [4]hading.muhaimin@ith.ac.id
[*)] Corresponding email

Abstrak—Penambangan data (*Data Mining*) merupakan teknik penting dalam era *Big Data* untuk menggali wawasan berharga dari kumpulan data besar. Tantangan utama dalam bidang ini adalah menjaga privasi individu, khususnya pada data *outlier* yang sensitif yang mengandung informasi pribadi. Penelitian ini bertujuan untuk membandingkan efektivitas algoritma *clustering* PAM, CLARA, CLARANS, dan ECLARANS dalam mendeteksi outlier serta mengevaluasi perlindungan privasi menggunakan metode Gaussian Perturbation Random. Penelitian dilakukan menggunakan dua dataset kesehatan: Dataset Diabetes dari National Institute of Diabetes and Digestive and Kidney Diseases dan Dataset Breast Cancer Wisconsin. Hasil menunjukkan bahwa algoritma CLARA mendeteksi jumlah *outlier* terbanyak pada dataset besar, sementara ECLARANS menunjukkan efisiensi waktu terbaik pada dataset tertentu. Metode Gaussian Perturbation Random terbukti efektif dalam melindungi privasi outlier tanpa mengurangi akurasi deteksi. Kesimpulannya, CLARA merupakan algoritma yang paling menjanjikan untuk mendeteksi outlier sambil menjaga privasi data, berkat pendekatan sampling yang efisien. Temuan ini memberikan kontribusi penting dalam penerapan *data mining* yang aman dan privasi yang terjaga, khususnya dalam domain data kesehatan.

Kata Kunci: Penambangan Data, Pengelompokan, Perlindungan Privasi, Deteksi Outlier, Gaussian Perturbation Random

Abstract—Data mining is a crucial technique in the era of Big Data for extracting valuable insights from large datasets. A major challenge in this field is ensuring individual privacy, particularly for sensitive outlier data that may contain personal information. This study aims to compare the effectiveness of clustering algorithms—PAM, CLARA, CLARANS, and ECLARANS—in detecting outliers and to evaluate privacy protection using the Gaussian Perturbation Random method. The research utilized two health datasets: the Diabetes Dataset from the National Institute of Diabetes and Digestive and Kidney Diseases and the Breast Cancer Wisconsin Dataset. The results indicate that the CLARA algorithm identified the highest number of outliers in large datasets, while ECLARANS demonstrated the best time efficiency for certain datasets. The Gaussian Perturbation Random method proved effective in protecting the privacy of outliers without compromising detection accuracy. In conclusion, CLARA is the most promising algorithm for outlier detection while preserving data privacy, owing to its efficient sampling approach. These findings provide significant contributions to the implementation of secure and privacy-preserving data mining, particularly in the domain of health data.

Keywords: Data Mining, Clustering, Privacy Preserving, Outlier Detection, Gaussian Perturbation Random.

1. INTRODUCTION

Data mining is the process of extracting predictive information from extensive databases (Big Data) and is a formidable technology with significant potential for analyzing critical information within its data warehouse [1]. Data mining forecasts future trends and behaviors, allowing businesses to make proactive and educated decisions based on data and statistics. The demand for users to gather and utilize extensive data has increased significantly. With the emergence of computers and extensive digital storage systems, users commenced the collection and storage of diverse data, leveraging computational capabilities to organize this amalgamation of information. Regrettably, enormous data sets were organized in disparate ways, which led to the development of structured databases and database management systems.

An efficient database management system is a crucial resource for managing big data, particularly for the effective retrieval of specific information from extensive data sets when required. The expansion of database management technologies has significantly facilitated the extensive accumulation of data. Currently, consumers are capable of managing greater volumes of information from corporate transactions and scientific data, including satellite imagery, textual reports, and military intelligence. Information retrieval is inadequate for decision-making due to the extensive data sets that generate new requirements for improved managerial choices. These requirements include automatic data summarization, extraction of the core essence of stored information, and identification of patterns in raw data, hence employing data mining to evaluate and extract information from extensive databases.

Privacy is defined as the Preserving of an individual's information. The Preserving of privacy has emerged as a significant concern in data mining research. The fundamental necessity of privacy-preserving data

mining is to safeguard the input data while enabling the data miner to derive valuable knowledge models [2]. Several privacy-preserving data mining techniques have recently been introduced utilizing cryptographic or statistical methodologies. Cryptographic methods guarantee robust privacy and precision via safe multi-party computation, although generally exhibit suboptimal performance. Statistical methodologies have been employed for decision trees, association rules, and clustering, gaining popularity mostly due to their superior performance.

A conventional dictionary definition of privacy concerning data is “freedom from unauthorized intrusion” [3]. When the user has granted permission for data use in a specific data mining endeavor, privacy concerns are eliminated. Nonetheless, if the user lacks authorization, it results in a breach of privacy. Privacy pertains to "individually identifiable information." Various techniques, including randomization, k-anonymity, distributed privacy preservation, query auditing, data posting, and cryptographic methods, have been proposed in recent years for executing privacy-preserving data mining. Privacy-preserving data mining strategies must guarantee that any disclosed information is protected [4]. Any data that fails to furnish complete and accurate information regarding a specific individual qualifies as privacy data. Conversely, the proliferation of information regarding an individual may be deemed a nuisance that could hinder data mining efforts, given the objective is to produce insights. While the focus is frequently on a collective, acquiring additional information about the group enhances understanding of the people inside it. This necessitates measuring both the acquired data and the capacity to associate it with a specific individual.

Although privacy in data mining remains inadequately regulated, numerous applications demonstrate that privacy-preserving data mining can yield valuable insights while adhering to established privacy protection norms. The issue of privacy-preserving data mining has gained prominence in recent years owing to the enhanced capacity to store personal user data and the growing sophistication of data mining algorithms that exploit this information [5].

This work aims to compare the effectiveness of four clustering algorithms in detecting sensitive outliers and deploying the use of Gaussian Random Perturbation in maintaining health data privacy. The results of this research are expected to provide practical solutions to address privacy protection in data mining in the health domain.

2. RESEARCH METHODS

Certain clustering methods, like PAM (Partitioning Around Medoid), CLARA (Clustering Large Applications), CLARANS (Clustering Large Applications Based On Randomized Search), and ECLARANS(Enhanced Clarans) can manage outliers; nevertheless, their primary focus is on identifying clusters, with outliers frequently seen as noise inside the clustering framework. Typically, outliers are either disregarded or accepted throughout the clustering process, as these algorithms are designed to generate significant clusters, hence hindering their efficacy in outlier detection [6].

The majority of outlier detection techniques are grounded in statistical principles. These approaches can be categorized into two types: distribution-based and depth-based. Distribution-based approaches employ a standard distribution to model the dataset. Outliers are identified according to a probability distribution. The primary issue encountered by distribution-based approaches is the presumption that the underlying data distribution is predetermined. Nevertheless, in several situations, preceding knowledge is not always accessible, and the endeavor to conform the data to a standard distribution is considerable. Distance-based methods classify a point as an outlier if its vicinity encompasses less than pct% of the complete dataset. This notion generalizes numerous principles from distribution-based methods and exhibits superior computational complexity. Long-distance strategies are implemented to enhance detection efficacy [6], [7].

Deviation-based approaches detect outliers by analyzing the primary attributes of objects within a collection, categorizing those that diverge from these characteristics as outliers. It employs a "local outlier factor" (LOF) to assess the likelihood of an object being an outlier. The LOF value of an object is derived by a comparison of its density with that of its environment, hence providing superior modeling capabilities compared to distance-based methods, which rely solely on the object's own density [8].

The primary aim of this research is to identify outliers through the application of clustering methods. The identified outliers are classified as sensitive outliers. Preserving sensitive outliers through a privacy strategy that involves altering the data pieces within the dataset. Subsequent to change, the identical clustering approach is employed for outlier detection. Proceed to ascertain the presence of outliers. The efficacy of the clustering algorithm for outlier detection and the privacy technique is evaluated.

2.1 Methodology

Figure 1 outlines the research methodology in a structured sequence, beginning with input data preparation, followed by preprocessing, outlier detection using clustering algorithms (PAM, CLARA, CLARANS, and ECLARANS) application of privacy protection methods, and concluding with evaluation metrics. This framework ensures consistency in detecting sensitive outliers while preserving data privacy.

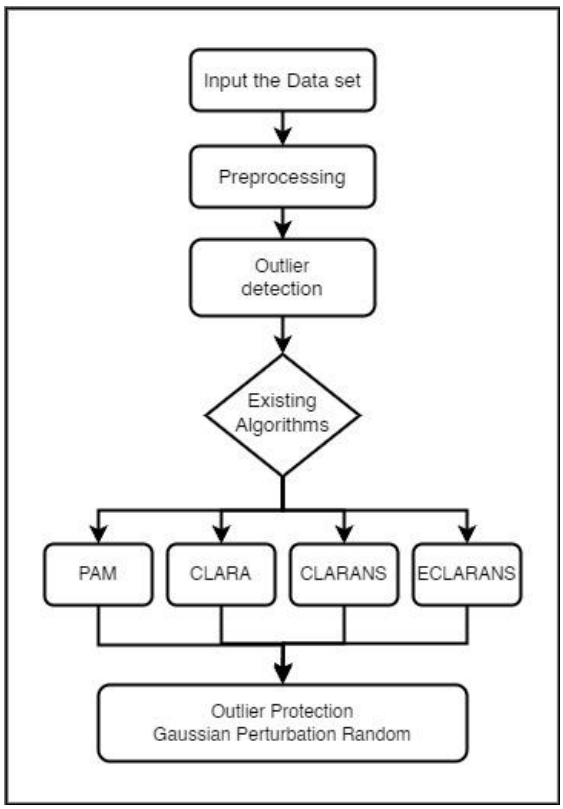

Figure 1. Methodology

2.2 Utilised Dataset

The Diabetes dataset from the National Institute of Diabetes and Digestive and Kidney Diseases [9] and the Breast Cancer Wisconsin dataset [10] are utilised for outlier detection and safeguarding. The datasets are sourced from Kaggle (https://www.kaggle.com), a platform that facilitates learning and sharing among data scientists and machine learning practitioners. The diabetes dataset contains 768 records with 9 attributes Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome). The Breast Cancer dataset contains 569 records and 12 attributes (id, diagnosis, radius_mean, texture_mean, perimeter_mean, area_mean, smoothness_mean, compactness_mean, concavity_mean, concave points_mean, symmetry_mean, fractal_dimension_mean). The attributes of these two datasets are multivariate and consist of integers (numeric).

Both datasets share the feature of being multivariate with diverse data distributions, ensuring that the clustering algorithms are tested across various levels of complexity. Their focus on healthcare data makes them ideal for exploring the balance between accurate outlier detection and effective privacy protection in sensitive domains.

2.3 Preprocessing

Data cleaning is a method for identifying and eliminating or rectifying faulty or erroneous entries from a dataset, table, or database, often referred to as data cleansing. Data cleansing largely pertains to databases and involves the detection of incomplete, erroneous, inaccurate, irrelevant, and missing data that may be changed, modified, or destroyed [11].

Data cleansing entails eliminating typographical errors and verifying values against a dataset of recognized entities. Post-cleansing, the dataset will exhibit consistency with the other datasets amalgamated from distinct databases. Advanced software solutions exist to cleanse dataset information utilizing algorithms and regulations. This study use the K-nearest neighbor imputation technique to address missing values by utilizing the average.

Preprocessing is a critical step to ensure the quality and consistency of the datasets before applying clustering algorithms. In this study, missing values were identified in both datasets, particularly in attributes such as glucose levels and BMI in the Diabetes dataset, which are common in medical records. The missing values were categorized as either missing completely at random (MCAR) or missing at random (MAR), depending on whether they were influenced by other variables. To handle these missing values, the K-nearest neighbor (KNN) imputation method was employed. This approach was chosen due to its ability to preserve the relationships between attributes by imputing values based on the similarity to neighboring data points. It is particularly suitable for multivariate datasets, as it ensures that the imputed values are consistent with the overall data distribution. Unlike mean or median imputation, KNN accounts for the variability within the dataset, reducing the risk of introducing bias.

In addition to imputing missing values, further data cleaning steps were undertaken to address potential outliers or inconsistencies in the data. Typographical errors and extreme outliers, which might skew clustering results, were corrected or removed to maintain data integrity. By employing the KNN method for imputation and rigorous data cleaning procedures, the datasets were prepared to ensure accurate and reliable results during outlier detection and privacy protection analysis.

2.4 Approaches to Outlier Detection

Outlier detection is a significant data mining task, known as outlier mining. Outliers are entities that deviate from the typical patterns of the data. This works using 4 clustering techniques shown in the table 1.

Table 1. Comparison of clustering techniques

Techniques	Approach
PAM	Uses k-medoids to form clusters based on the most representative objects (medoids).
CLARA	Uses a sampling technique to select a subset of data, which is then clustered using PAM.
CLARANS	Selects medoids randomly and performs local searches to find the optimal configuration
ECLARANS	Optimizes the random node selection process in CLARANS to reduce the required number of iterations.

2.4.1 PAM (Partitioning Around Medoid)

PAM use the k-medoids algorithm for clustering. It is significantly more resilient than k-means in the face of noise and outliers. It comprises two phases: the Build phase and the Swap phase [6]. Build phase: This stage systematically picks k objects positioned centrally. These k items serve as k-medoids. Swap phase: Calculates the aggregate cost for each combination of picked and unselected objects. The following PAM techniques are presented in Table 2.

Table 2. PAM Procedures

Sequence	Activity
1	Insert the dataset D
2	Select k objects randomly from the dataset D
3	Compute the total cost T for each picked object Si and non-selected object S_h
4	For each pair, if T $S_i < 0$, then replace it with S_h
5	Identify the analogous medoid for each unselected object
6	Iterate stages 2, 3, and 4 until the medoids are identified

2.4.2 CLARA (Clustering Large Applications)

CLARA is implemented to address the issue of PAM. This is effective with larger datasets than PAM. This strategy utilizes only a subset of data from the entire dataset rather than the complete dataset. The PAM algorithm randomly selects data and identifies the medoid[6]. The following CLARA techniques are presented in Table 3.

Table 3. CLARA Procedures

Sequence	Activity
1	Insert the dataset D
2	Repeat n times (e.g., five times)
3	Randomly select a sample S from the dataset D
4	Execute the PAM algorithm on the sample S to derive the medoids M
5	Classify the complete dataset D into Cost1 through Costk
6	Compute the average dissimilarity for the clusters generated in the complete dataset

2.4.3 CLARANS (Clustering Large Applications Based On Randomized Search)

This approach resembles PAM and CLARA. Clarans commences with the choosing of random medoids. It dynamically attracts neighbors. It verifies the "max neighbor" for the purpose of swapping. If the pair is negative, it selects an alternative set of medoids. Alternatively, it selects the existing medoid configuration as the local optimum and initiates a fresh random medoid selection. It suspends the process until it identifies the optimal option [12]. The following CLARANS techniques are presented in Table 4.

Table 4. CLARANS Procedures

Sequence	Activity
1	Specify the input parameters numlocal and maxneighbour
2	Select k objects randomly from the database object D
3	Designate these K items as selected S_i and the others as non-selected S_h.
4	Determine the cost T for the chosen S_i
5	If T is negative, update the medoid set. Alternatively, the selected medoid is designated as the local optimum.
6	Initiate the selection of an alternative medoid set and identify another local optimum
7	CLARANS ceases operation until it yields the optimal result

2.4.4 ECLARANS (Enhanced Clarans)

This approach differs from PAM, CLARA, and CLARANS. This strategy can enhance the precision of outlier detection. ECLARANS is an innovative partition algorithm that enhances the CLARANS form cluster by strategically selecting the appropriate arbitrary node rather than employing a random search method. This algorithm resembles CLARANS, except the selection of an arbitrary node decreases the number of CLARANS iterations [12]. The following ECLARANS techniques are presented in Table 5.

Table 5. ECLARANS Procedures

Sequence	Activity
1	Specify the input parameters numlocal and maxneighbour. Set i to 1 and mincost to a substantial value.
2	Calculate the distance between each data point.
3	Select n data points exhibiting the greatest distance.
4	Assign the current position to a specified node in n:k
5	Assign the value of 1 to j.
6	Evaluate a random neighbor S of the current node and, based on point 6, calculate the cost differential between the two nodes.
7	If S presents a reduced cost, designate the current as S and proceed to Step 5.
8	Alternatively, increment j by 1. If j is the maximum neighbor, proceed to Step 6.
9	If j exceeds maxneighbour, assess the current cost against the minimal cost. If the former is inferior to the small cost, assign the current cost to the small cost and designate the current as the best node.
10	Increment i by 1. If i exceeds numlocal, output the optimal node and terminate. If not, proceed to Step 4.

2.4.5 Privacy Techniques Based on Gaussian Perturbation Random Method

Data perturbation is a method of privacy-preserving data mining applied to electronic health records (EHR). Two primary forms of data disruption are suitable for the protection of electronic health record (EHR) data. The initial type is referred to as the probability distribution approach, while the subsequent type is designated as the value distortion approach. Data perturbation is regarded as a straightforward and efficient method for safeguarding sensitive electronic information from unauthorised access. Data perturbation is seen as a superior data protection method in healthcare compared to deidentification/re-identification, owing to the increased probability of an attack that associates a public dataset with an original identifier or subject. Data disruption is seen as a more effective method for enhancing EHR security [13].

The probability distribution method utilises data by substituting it with samples from the identical distribution or from the distribution itself. The value distortion method modifies the data using multiplicative or additive perturbations, or other stochastic processes. This is deemed more effective than the prior form of disturbance. This method constructs a decision tree classifier in which each component is subjected to a random perturbation derived from a Gaussian distribution, for instance. Data mining is reconstructing the original data distribution from its altered version. Critics, however, highlight that random additive noise can be eliminated, potentially leading to violations of EHR privacy [2].

Following the outlier detection process, we regard the identified outliers as sensitive information requiring protection. The subsequent stage is to provide a method for safeguarding the privacy of sensitive information. We offer an innovative method of privacy engineering utilising the Gaussian Perturbation Random Method. The primary aim of this proposed method is to apply a rounding technique to outlier data, hence safeguarding the information. This method initialises their ensemble for the improved CLARANS data representation.

For instance, consider an outlier in a diabetes dataset where a patient has an unusually high blood glucose level of 300 mg/dL. Without protection, this data point could be linked back to an individual, posing a privacy risk. By applying Gaussian Perturbation Random, the glucose value may be slightly altered, such as being changed to 290 mg/dL or 310 mg/dL, depending on the noise level introduced. This ensures that the overall data distribution remains intact for analysis while preventing the exact value from being traced to a specific individual. [14]

This approach effectively prevents re-identification of individuals based on outlier values, ensuring that sensitive health information remains protected. By maintaining the balance between data utility and privacy, the Gaussian Perturbation Random method provides a robust solution for privacy-preserving data mining in sensitive domains such as healthcare [14]. The perturbation data derived from the Gaussian Perturbation Random Method is presented in Table 6 as follows.

Table 6. Data perturbation by Gaussian Perturbation Random Method

Sequence	Activity
Input : Dataset Object	
Output : Perturbed dataset	
Method :	
1	Take into account sensitive outliers from the dataset.
2	Select clusters one by one
2.1	Consider data items
2.2	Consolidate the proximate outliers and group the data elements
2.3	Determine the mean value of outliers and the mean value of clustered data pieces.
2.4	Determine the disparity between the mean of outliers and the mean of clustered data points.
2.5	Randomly choose outliers and modify their values by either augmenting or diminishing their deviations.
3	Execute the identical procedure for all clusters.

3. RESULTS AND DISCUSSIONS

This research was conducted in Jupyter utilising Python 3. Diabetes dataset from the National Institute of Diabetes and Digestive and Kidney Diseases and Breast Cancer Wisconsin dataset for the purposes of outlier discovery and prevention.

3.1 Accuracy of Outliers

The accuracy of outlier detection is assessed by comparing the number of true outliers detected by the algorithms with the actual outliers in the dataset detection is assessed to determine the quantity of outliers identified by the PAM, CLARA, CLARANS, and ECLARANS clustering algorithms.

Table 7. Outliers Detected.

Dataset	PAM	CLARA	CLARANS	ECLARANS
Diabetes	39	65	39	39
Breast Cancer	29	35	29	29

In the diabetic dataset, PAM, CLARANS, and ECLARANS identified 39 outliers, while CLARA discovered 65 outliers. The same methods employed for outlier detection in the Breast Cancer dataset are also included in Table 6. Consequently, it can be demonstrated that the CLARA algorithm enhances the precision of outlier detection.

CLARA is capable of detecting more outliers primarily due to its sampling-based approach. Unlike PAM, which analyzes the entire dataset directly, CLARA selects multiple subsets of the data and applies clustering algorithms to these smaller samples. This method increases the likelihood of identifying variations and anomalies within different segments of the dataset, leading to the detection of additional outliers. Additionally, CLARA reduces computational complexity by focusing on representative subsets rather than processing the entire dataset, making it more efficient for handling large-scale data while maintaining high sensitivity to outliers.

.

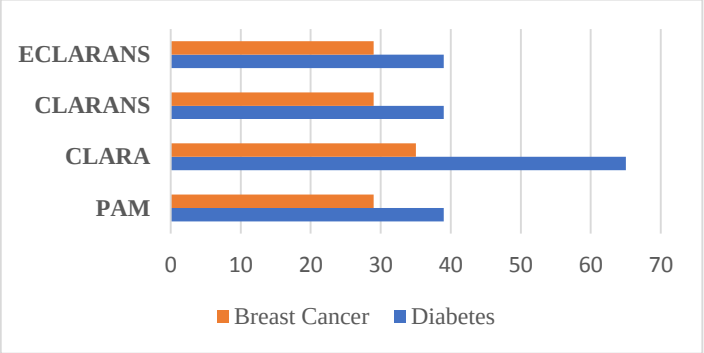

Figure 2. Outlier Accuracy

The graph illustrates the quantity of outliers identified by the current clustering algorithms: PAM, CLARA, CLARANS, and ECLARANS. The CLARA clustering algorithm has identified a greater proportion of outliers than other algorithms. These findings highlight that CLARA's sampling strategy enhances its effectiveness in identifying anomalies, especially in large datasets where outliers may be sparsely distributed. The results demonstrate the advantages of using sampling techniques in clustering-based outlier detection while preserving computational efficiency.

3.2 Time Complexity of the Clustering Algorithm

The time complexity is measured by recording the execution time of each algorithm for identifying outliers on the given dataset is evaluated based on the duration needed to identify outliers using the PAM, CLARA, CLARANS, and ECLARANS clustering algorithms.

Table 8. Time Complexity (in seconds)

Dataset	PAM	CLARA	CLARANS	ECLARANS
Diabetes	0.41916	0.97229	0.62296	0.45726
Breast Cancer	0.46169	0.14628	0.82911	0.88481

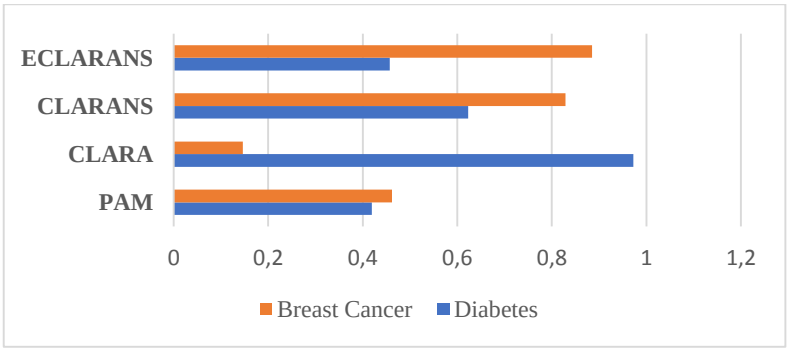

Figure 3. Time Complexity (in seconds)

The comparative analysis of the algorithms' time complexity reveals that the CLARA algorithm exhibits superior efficiency on the breast cancer dataset, Although CLARA takes longer to identify outliers, its ability to detect more outliers justifies the additional time required.

3.3 Outlier Protection Results

Gaussian Perturbation Random Method is a privacy-preserving technique designed to protect sensitive outliers by introducing controlled noise into the data. This method applies a Gaussian distribution to perturb outlier values, ensuring that individual records cannot be directly linked back to their original values while maintaining the overall data structure. Unlike traditional anonymization techniques that may remove or obscure data entirely, Gaussian perturbation modifies the values slightly in a way that preserves statistical properties, allowing meaningful analysis to continue. This approach is particularly beneficial in healthcare datasets, where maintaining data integrity while ensuring patient privacy is crucial [14].

Once outliers were detected using clustering algorithms (PAM, CLARA, CLARANS, and ECLARANS), they were considered sensitive information requiring protection. The Gaussian Perturbation Random Method was then applied to modify the values of these outliers, adding noise based on a Gaussian distribution. This transformation

ensured that the original values were obscured without distorting the broader trends in the dataset, preserving the usability of the data for further analysis.

The method successfully reduced the risk of re-identification by modifying outlier values while maintaining the overall distribution of the data. Even after perturbation, the general pattern of the dataset remained intact, ensuring that analytical results remained valid. Graphical results demonstrated that the distribution of outliers, after applying perturbation, still closely resembled the original dataset, confirming the method's effectiveness.

The effectiveness of this protection technique was further analyzed by comparing the accuracy of clustering results before and after perturbation. The findings revealed that Gaussian Perturbation Random not only safeguarded sensitive outliers but also preserved the ability to detect anomalous patterns within the dataset. By balancing privacy protection with data utility, this method provides a robust solution for privacy-preserving data mining, especially in sensitive domains such as healthcare.

The Gaussian Perturbation Random Method is a rounding technique employed to mitigate outlier information, therefore safeguarding the data. This approach initializes the ensemble for data representation of the four utilized methods.

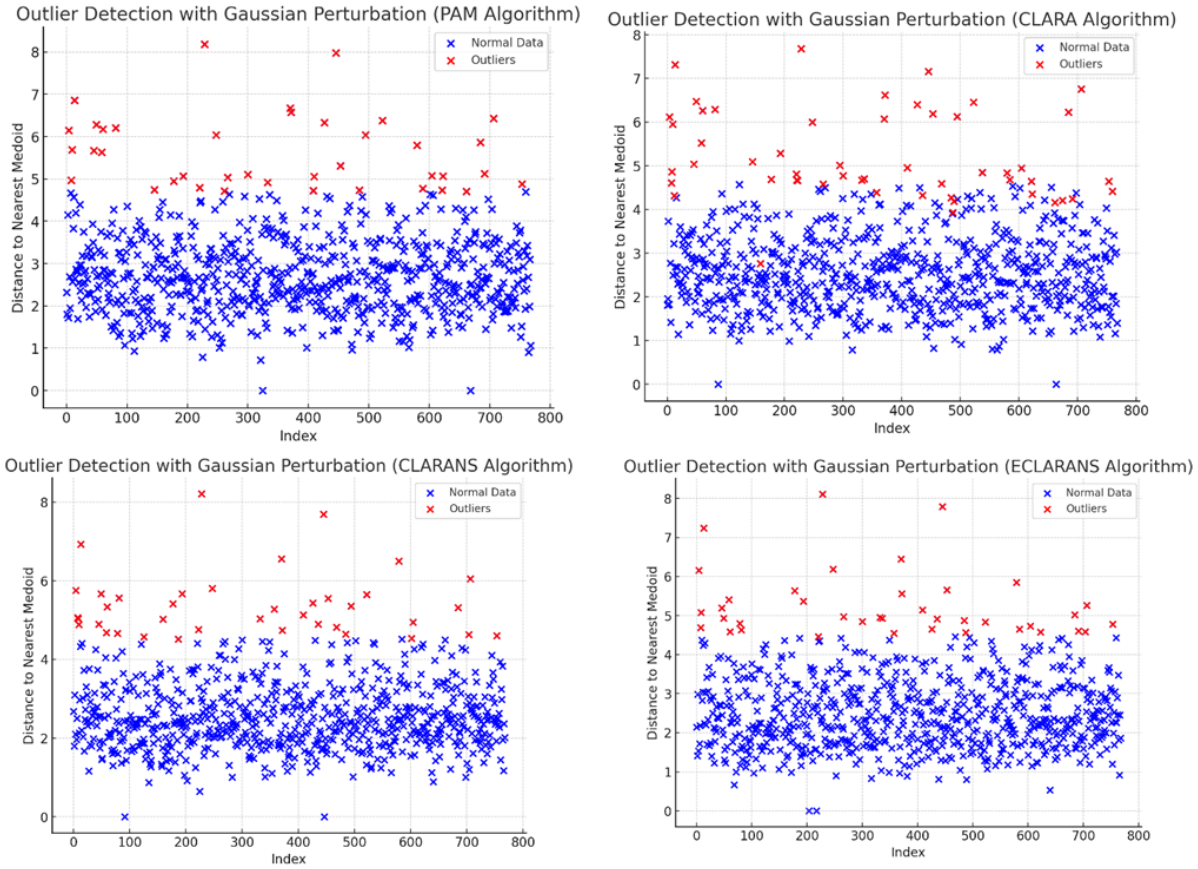

Figure 4. Gaussian Perturbation Random outlier detection utilizing four methods on the diabetes dataset

Figure 4 displays the results of outlier detection using the Gaussian Perturbation Random technique. This method is applied to the PAM, CLARA, CLARANS, and ECLARANS algorithms on the diabetes dataset. The red dots (outliers) are positioned further from the cluster center (medoid) at the top of the graph, while the blue dots (normal data) are closer to the medoid and clustered towards the bottom of the graph. PAM Algorithm: Red dots (outliers) are positioned at the top, whilst blue dots (normal data) are predominantly clustered below, signifying a reduced distance to the medoid. The CLARA algorithm identifies a greater number of outliers compared to PAM, exhibiting a similar distribution; however, the overall distance pattern appears marginally more dispersed. The CLARANS algorithm exhibits an outlier distribution akin to PAM; however, certain outliers are observed at greater distances, suggesting discrepancies in data clustering by CLARANS. The ECLARANS Algorithm exhibits an outlier distribution akin to CLARANS, revealing several outliers that are more distant from the cluster centroid.

The picture displays the results of outlier detection employing the Gaussian Perturbation Random technique across four clustering algorithms: PAM, CLARA, CLARANS, and ECLARANS. Each figure illustrates the correlation between each data point and its nearest medoid, with blue dots denoting normal data and red dots indicating outliers. In each approach, the distribution pattern of normal data and outliers appears consistent, with

outliers positioned at a greater distance from the cluster's center, signifying distinct characteristics relative to normal data.

In the PAM algorithm, outliers are distinctly positioned at the upper section of the graph both with and without Gaussian noise. This indicates that despite the introduction of Gaussian noise, the algorithm effectively identifies outliers, with the quantity of recognized outliers remaining consistent with the condition devoid of noise. In contrast, with CLARA, the distribution of outliers appears more dispersed after noise is introduced compared to the more compact distribution observed without noise. This suggests that the employed sampling strategy can yield variances in outlier identification outcomes, making CLARA more sensitive to perturbations, while PAM maintains stability in detecting outliers regardless of noise application.

The CLARANS and ECLARANS algorithms yield a pattern akin to PAM, albeit with minor differences in the quantity and proximity of outliers identified. This illustrates the intricate nature of the clustering technique, wherein a more dynamic medoid replacement process may produce varying outcomes.

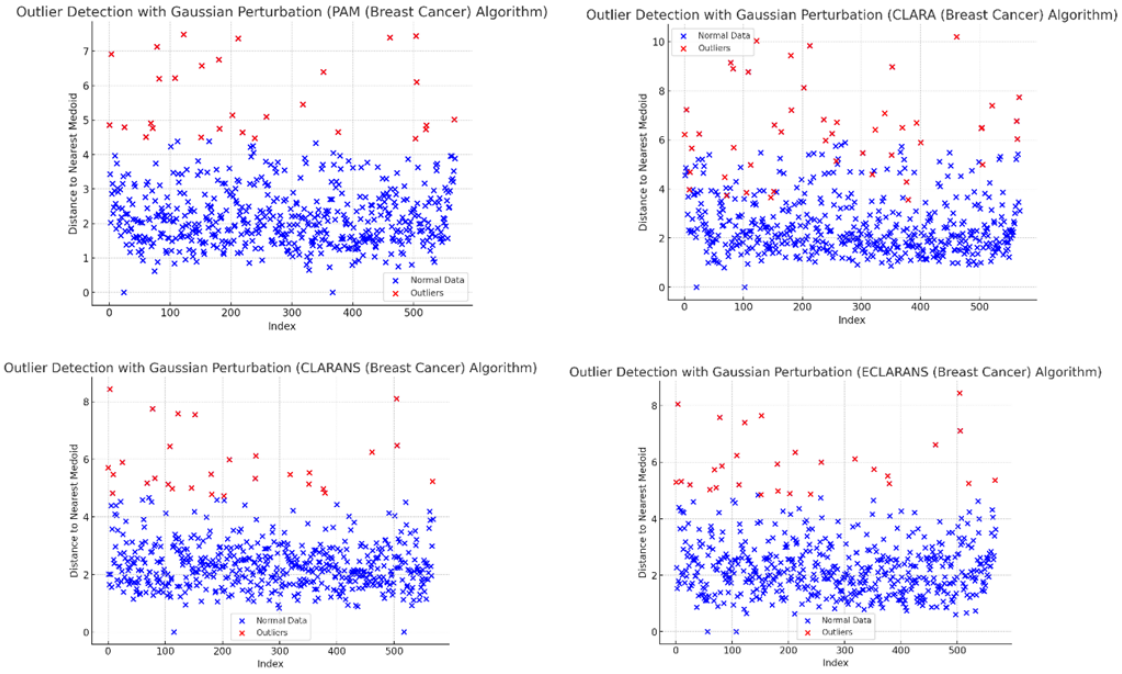

Figure 5. Gaussian Perturbation Random outlier detection utilizing four methods on the breast cancer dataset

Figure 5 illustrates the outcomes of outlier detection employing the Gaussian Perturbation Random approach on the PAM, CLARA, CLARANS, and ECLARANS algorithms utilizing the breast cancer dataset: PAM (Breast Cancer): The red dots in this picture signify the identified outliers. Despite the addition of Gaussian noise, the distribution of outliers is predominantly focused at the upper section of the graph, suggesting that the algorithm can identify data with varying features, even in the presence of minimal noise. CLARA (Breast Cancer): This illustration presents the outcomes of outlier detection with the CLARA algorithm. The quantity of identified outliers seems to be elevated, since numerous red dots signify that certain data points are distanced from the typical cluster. This illustrates CLARA's proficiency in managing data sampling while efficiently identifying outliers. CLARANS (Breast Cancer): The findings for the CLARANS algorithm exhibit a distribution analogous to those of PAM. Nonetheless, there exists variability in the location of the identified outliers, indicating the more intricate dynamics of the medoid inside this approach. ECLARANS (Breast Cancer): The depiction for ECLARANS exhibits an analogous outlier distribution to that of CLARANS. Despite the detection of outliers, the majority of the data is concentrated around the lower values, demonstrating the resilience of this method against perturbations.

The analysis of outlier detection results utilising the PAM, CLARA, CLARANS, and ECLARANS algorithms on diabetes and breast cancer datasets indicates that the Gaussian Perturbation Random method is effective for identifying outliers while preserving individual privacy. In all datasets, despite the introduction of Gaussian perturbation, the system effectively identified outliers, demonstrating consistency in both the quantity and distribution of discovered anomalies. Despite adding noise, the Gaussian Perturbation Random method effectively identifies outliers while ensuring that the original dataset's privacy is preserved, with minimal distortion in outlier detection results.

The diabetes dataset exhibited no variation in the number of detected outliers following the application of perturbation, demonstrating the algorithm's robustness to minor noise. The breast cancer dataset exhibited variability in the amount of outliers identified by each algorithm, indicating discrepancies in the data clustering methodology. This indicates that while perturbation may conceal information, the algorithm remains effective in identifying anomalous data.

Maintaining privacy is crucial, particularly for sensitive data like health information. Utilising Gaussian perturbation safeguards personal information potentially included in outlier data, hence mitigating the risk of de-anonymization and privacy infringements. This strategy safeguards individual confidentiality during data analysis, enabling academics and practitioners to employ data for research and analysis without revealing sensitive information. In the contemporary digital era, where privacy is a significant issue, the application of this technique is crucial for preserving integrity and confidence in data utilisation.

4. CONCLUSION

The CLARA algorithm demonstrates the most potential for preserving privacy in the analysed diabetic and breast cancer datasets. This results from CLARA's capacity to effectively sample and cluster extensive data subsets, together with its proficiency in identifying outliers without compromising the overall analytical accuracy. This method mitigates the risk of revealing sensitive information by recognising anomalies in the sampling process, particularly concerning data related to an individual's health.

A primary advantage of CLARA is its methodology of sampling from the complete dataset, ensuring that not all of an individual's data is revealed throughout the study. This offers an extra degree of privacy protection, as sensitive information from a person is not consistently apparent in the results. Consequently, while CLARA may identify a greater number of outliers, this sampling method mitigates the risk of de-anonymization and offers enhanced privacy for the individuals whose data are examined.

The PAM, CLARANS, and ECLARANS algorithms demonstrate the capability to identify outliers, albeit using distinct methodologies. PAM demonstrates commendable efficacy in outlier detection accuracy; nevertheless, it does not employ sampling, resulting in the direct inclusion of each individual in the dataset in the study. This may elevate the risk of revealing sensitive information, as the identified outliers can exhibit distinctive traits of persons, rendering them more vulnerable to further investigation.

The variation in outlier detection results among these algorithms can be attributed to their underlying methodologies. CLARA's reliance on sampling allows it to uncover more outliers, but this comes with the trade-off of potential sampling bias. PAM, being a k-medoids-based approach, offers high accuracy for small datasets but struggles with larger datasets due to its exhaustive computation. CLARANS introduces a more dynamic approach with randomized search, which increases efficiency but may lead to inconsistent outlier detection across different runs. ECLARANS improves upon CLARANS by refining node selection, ensuring better stability and accuracy, especially for large and complex datasets. These differences highlight the need for selecting an appropriate algorithm based on dataset characteristics and computational constraints.

In summary, while each method possesses distinct advantages and limitations regarding privacy, CLARA is superior for safeguarding privacy in both datasets. The employed sampling methodology offers adaptability and enhanced safeguarding for sensitive information, rendering it a proficient option for data analysis in an era that progressively prioritises individual privacy protection. This underscores the significance of evaluating privacy considerations when selecting an algorithm for data analysis, particularly in domains requiring sensitive information like healthcare.

REFERENCES

[1] M. K. Gupta and P. Chandra, “A comprehensive survey of data mining,” *International Journal of Information Technology*, vol. 12, no. 4, pp. 1243–1257, Dec. 2020, doi: 10.1007/s41870-020-00427-7.

[2] A. Pika, M. T. Wynn, S. Budiono, A. H. M. ter Hofstede, W. M. P. van der Aalst, and H. A. Reijers, “Privacy-Preserving Process Mining in Healthcare,” *Int J Environ Res Public Health*, vol. 17, no. 5, p. 1612, Mar. 2020, doi: 10.3390/ijerph17051612.

[3] J. Dong, A. Roth, and W. J. Su, “Gaussian Differential Privacy,” *J R Stat Soc Series B Stat Methodol*, vol. 84, no. 1, pp. 3–37, Feb. 2022, doi: 10.1111/rssb.12454.

[4] M. A. P. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil, “Efficient privacy preservation of big data for accurate data mining,” *Inf Sci (N Y)*, vol. 527, pp. 420–443, Jul. 2023, doi: 10.1016/j.ins.2019.05.053.

[5] V. S. Naresh and M. Thamarai, "Privacy- preserving data mining and machine learning in healthcare: Applications, challenges, and solutions," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, Mar. 2023, doi: 10.1002/widm.1490.

[6] J. Alvariño-Durán, B. Hernández-Ocaña, J. Hernández-Torruco, and O. Chávez-Bosquez, "Detection of Cardiac Arrhythmias Using Unsupervised Learning: A Preliminary Approach Based on PAM and CLARA Clustering Algorithms," 2024, pp. 594–601. doi: 10.1007/978-3-031-62502-2_67.

[7] B. Dastjerdy, A. Saeidi, and S. Heidarzadeh, "Review of Applicable Outlier Detection Methods to Treat Geomechanical Data," *Geotechnics*, vol. 3, no. 2, pp. 375–396, May 2023, doi: 10.3390/geotechnics3020022.

[8] X. Du, E. Zuo, Z. Chu, Z. He, and J. Yu, "Fluctuation-based outlier detection," *Sci Rep*, vol. 13, no. 1, p. 2408, Feb. 2023, doi: 10.1038/s41598-023-29549-1.

[9] Mehmet Akturk, "Diabetes Dataset : This dataset is originally from the N. Inst. of Diabetes & Diges. & Kidney Dis.," Kaggle.

[10] UCI Machine Learning, "Breast Cancer Wisconsin (Diagnostic) Data Set : Predict whether the cancer is benign or malignant," Kaggle.

[11] S. E. Whang, Y. Roh, H. Song, and J.-G. Lee, "Data collection and quality challenges in deep learning: a data-centric AI perspective," *The VLDB Journal*, vol. 32, no. 4, pp. 791–813, Jul. 2023, doi: 10.1007/s00778-022-00775-9.

[12] P. Sarang, "CLARANS," 2023, pp. 237–242. doi: 10.1007/978-3-031-02363-7_14.

[13] S. Turgay and İ. İlker, "Perturbation Methods for Protecting Data Privacy: A Review of Techniques and Applications," *Automation and Machine Learning*, vol. 4, no. 2, 2023, doi: 10.23977/autml.2023.040205.

[14] J. Zhou, W. Lan, and H. Wang, "Asymptotic covariance estimation by Gaussian random perturbation," *Comput Stat Data Anal*, vol. 171, p. 107459, Jul. 2022, doi: 10.1016/j.csda.2022.107459.