SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

A Comparative Analysis Of Novel Privacy-Preserving Data Mining Techniques For Protecting Confidential And Sensitive Data Items

Dr.Vigneshwar Mahendran[1] , Dr. G. Karthikeyan[2] , R. Naveenkumar[3] , Dr.M.Subramaniakumar[4] , Dr. C Uma[5] , Dr.B.Senthilkumaran[6], Dr.A.Jeeva[7]

[1]Junior Resident, Department Of ENT, Saveetha Medical College And Hospital, SIMATS, Chennai, Tamil Nadu Mail Id : Vigneshvicky2996@Gmail.Com ORCID Id: 0009-0004-1580-1190
[2]ASSISTANT PROFESSOR, DEPARTMENT OF SOFTWARE SYSTEMS, PSG COLLEGE OF ARTS & SCIENCE,COIMBATORE - 641 014. Karthikeyan_G@Psgcas.Ac.In
[3]Dept Of CSE, School Of Engineering And Technology, CGC University Mohali-140307, Punjab India. Drnk1983@Gmail.Com 0000-0001-9033-9400
[4]Assistant Professor, Kristu Jayanti Institute Of Technology, KRISTU JAYANTI UNIVERSITY, Bengaluru. Email: Subramaniakumar.Sk@Gmail.Com
[5]Assistant Professor, Department Of Computer Technology And Information Technology, Kongu Arts And Science College, (Autonomous), Erode, Tamilnadu, India. Mail Id: Umachinnasamy@Gmail.Com 0009-0005-0097-6403
[6]Associate Professor, Department Of Computer Science And Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute Of Science And Technology Chennai. Skumaran.Gac16@Gmail.Com Official: Drsenthilkumaranb@Veltech.Edu.In 0000-0002-7111-1950
[7]Assistant Professor, Department Of Mathematics, Vel Tech Rangarajan Dr.Sagunthala R&D Institute Of Science And Technology, Avadi, Chennai-600062 Email Id:Drjeevaa@Veltech.Edu.In Orcid: 0000-0003-4087-2041

Abstract: Privacy-Preserving Data Mining (PPDM) deals with the issue of facilitating knowledge discovery whilst protecting the sensitive data in common datasets. This paper introduces three masking schemes of continuous numerical values MinMax, N-Point Crossover (NPC), and Hybrid Crossover masking schemes and compares it with the current Genetic Algorithm based Masking Technique (GAMT). The transformation model that was used to create masked microdata and maintain the statistical properties. It uses a synthetic worker- income data (3K-20K cases) in which income is considered as a confidential feature. Statistical accuracy, privacy protection accuracy and clustering accuracy of proposed Hybrid Crossover-based transformation are assessed with the help of five clustering algorithms: K-means clustering, X-means clustering, Filtered clustering, Expectation-Maximization (EM) and Make Density-Based Clustering (MDBC). The experiments show that the suggested Crossover-based transformation methods provide close statistical and privacy accuracy and high clustering consistency between original and masked databases. According to the findings, the proposed crossover-based masking schemes are effective to achieve the data utility and confidentiality hence suitable in the publication of secure data within the distributed data mining settings.

Keywords: Data Mining, Privacy, MinMax, N Point Crossover, GAMT, Clustering.

1. Introduction

The continuous and enormous digital data creation due to the use of enterprise systems, healthcare databases [1], financial record operations, and web-based systems has tremendously enhanced the need to have sophisticated analytical frameworks that can provide actionable knowledge. Data Mining (DM) enables the automatic identification of concealed patterns, correlations and structural regularities of large quantities of data. The important mining capabilities consist of classification, clustering, association rule mining, anomaly detection and dependency modelling [2]. The clustering is one of them which plays a central role in exploratory data analysis where the unclassified data is grouped into homogenous parts that are similar to each other in some measure. Clustering, in contrast to supervised learning, does not use available labeling of individuals into classes, and normally uses distance measures that meet metric space properties. To measure the intra-cluster

compactness and inter-cluster separation, cluster features are usually measured by centroid, radius, and diameter [3].

Although it has the benefits mentioned above, data mining raises substantial privacy issues where sensitive attributes are revealed in the process of data sharing or outsourcing. Organizations often have to publish datasets to third-party analysts or research communities but direct publication of microdata can result in confidentiality breaches [4]. Privacy-Preserving Data Mining (PPDM) has been created to overcome this difficulty and preserve sensitive information whilst maintaining the utility of data to perform analytical functions. Privacy preserving methods that are currently in use are such as randomization, k-anonymity, perturbation, sampling, as well as transformation-based masking. The major challenge is to produce a transformed dataset M01 of the original dataset Msuch that has similar statistical features and mining performance and the sensitive attributes are not revealed, as much as possible [5].

This study assumes continuous numerical properties as being kept confidential using transformation based masking schemes that take the form of M'=AMB+C wherein the transformations of matrices and regulated noise inclusion are executed to alter concealed values. Three masking strategies based on crossover and one of them, N-Point Crossover (NPC), are suggested and compared to currently used Genetic Algorithm based Masking Technique (GAMT). The efficacy of these methodologies is determined based on the accuracy of the statistical method, accuracy of privacy protection, and accuracy of clustering based on various algorithms of clustering methods. The test-based system shows that crossover-based masking methods are able to maintain the analytical consistency as well as provide a high level of confidentiality thus facilitating the publication of data in a secure way in distributed and collaborative mining systems.

2. Related works

Recent Privacy-Preserving Data Mining (PPDM) literature has examined a variety of approaches, such as synthetic data, cryptography, anonymization, and federated learning mechanisms, to preserve confidentiality in the data analysis process. Kostopoulos et al. have discussed, benchmarking methods of privacy-preserving synthetic data generation and compared both statistical models and deep generative models of generating synthetic student data that could be used to perform educational data mining tasks. Their empirical analysis showed that generative models can share the same statistics and minimise the risk of disclosure, thus permitting safe processing on data without revealing original documents [6].

Information exchange models that are aimed at secure analytics have become popular. Konda et al. offered an in-depth introduction to privacy-saving technologies that facilitate the exchange of data in the age of artificial intelligence and data science safety. Their publication talks about cryptography, anonymization, and differential privacy technologies that can be incorporated in the large-scale data ecosystem to guarantee confidentiality in the large knowledge discovery [7]. In the health care sector, Sakthidevi and Fathima came up with a threat intelligence model that incorporates privacy enhancing strategies to allow sharing of healthcare information in the cloud. Their structure integrates encryption, anomaly detection, and secure data exchange protocols to ensure unauthorized access and at the same time ensuring their analytics to be useful on medical datasets [8].

Distributed aggregation (only privacy preserving) has been explored in the context of distributed settings, including smart grids with IoT connections. The scheme of Lightweight Certificateless and Escrow-Free Privacy-Preserving Data Aggregation (LCE-PPDA) during the UAV-assisted smart grid systems was introduced by Chang et al. The protocol suggested is more efficient in computation and removes the overhead of handling certificates but ensures a high privacy level to distributed sensor data [9]. The concept of federated learning has become a promising paradigm of privacy-controlled collaborative learning. Zhou et al. introduced FairGFL, a federated learning framework focused on fairness, which ensures the privacy of the users and biasing in a multi-graph structure. They combine fairness constraints with graph-based federated optimization into their model which enhances both model equity and privacy preservation in distributed learning systems [10].

Classical data mining algorithms have also been processed using cryptographic secure computation techniques. Cooper et al. created cryptographic protocols that allow privacy keeping k-nearest neighbor (KNN) classification of outsourced data. Their approach will make sure that the training data and query inputs are encrypted throughout the computation, thus avoiding information leakage in the third-party cloud settings [11]. Moreover, dynamic privacy models have also been suggested to deal with new security threats. Priya and Kumari came up with quantum resilient privacy-preserving system using anonymization methods and adaptive security measures. They combine post-quantum cryptographic primitive and dynamical anonymization

techniques in order to increase resistance against future quantum-based threats in secure data mining processes [12].

3. Methodology - Hybrid Crossover-based transformation

The proposed Privacy-Preserving Data Mining (PPDM) framework is formally defined over a structured dataset $D \in \mathbb{R}^{n\times m}$, where each record $r_i = (a_1, a_2, \dots, a_m)$contains both categorical and numerical attributes. Let $A = \{a_1, a_2, \dots, a_m\}$denote the attribute space and $A_c \subseteq A$represent the subset of confidential continuous numerical attributes subject to masking. The objective is to construct a transformation function $T: \mathbb{R}^{n\times m} \to \mathbb{R}^{n\times m}$such that the transformed dataset $D' = T(D)$preserves analytical utility while minimizing disclosure risk. Formally,

$$D = \{r_1, r_2, \dots, r_n\}, r_i \in \mathbb{R}^m$$
$$A_c = \{a_j \mid a_j \in A,\ a_j \text{ is confidential and continuous}\}$$

The transformation process operates exclusively on the projection of Dover A_c, denoted as $X \in \mathbb{R}^{n\times k}$, where $k = |A_c|$. Thus,

$$X = \pi_{A_c}(D)$$

The transformed confidential data $X'$is generated using a composite transformation combining crossover-based perturbation and linear transformation. The general transformation model is expressed as

$$X' = AXB + C$$

where $A \in \mathbb{R}^{n\times n}$and $B \in \mathbb{R}^{k\times k}$are transformation matrices controlling structural distortion, and $C \in \mathbb{R}^{n\times k}$is a stochastic perturbation matrix. The transformation is constrained to preserve first-order statistical properties, ensuring

$$|\mu(X) - \mu(X')| \leq \epsilon$$

where

$$\mu(X) = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}$$

and $\epsilon$is a predefined tolerance threshold. Variance preservation is similarly constrained as

$$|\sigma^2(X) - \sigma^2(X')| \leq \delta$$

ensuring distributional similarity.

To model privacy protection formally, a probabilistic indistinguishability constraint is imposed. Let $X_{ij}$and X'_{ij}denote original and masked values, respectively. The privacy condition is defined as

$$P(X_{ij} = X'_{ij}) \approx 0$$

and extended to an adversarial inference setting as

$$I(X; X') \leq \gamma$$

where $I(\cdot)$denotes mutual information and $\gamma$is a minimal leakage threshold.

The masking mechanism is driven by crossover operations derived from evolutionary computation. Let each confidential value $x \in X$be represented as a digit sequence

$$x = (d_1, d_2, \dots, d_L), d_l \in \{0,1, \dots, 9\}$$

For two selected values $x_p$and x_q, an N-point crossover operation is defined as

$$x' = (d_1^{(p)}, \dots, d_c^{(p)}, d_{c+1}^{(q)}, \dots, d_L^{(q)})$$

where $c \in \{1,2, \dots, L-1\}$is the crossover point. The generalized multi-point crossover can be expressed as

$$x' = \sum_{l=1}^{L} \alpha_l d_l^{(p)} + (1-\alpha_l) d_l^{(q)}$$

where $\alpha_l \in \{0,1\}$defines the crossover mask.

For the MinMax strategy, let

$$x_{min} = \min(X), x_{max} = \max(X)$$

The crossover transformation is defined as

$$x' = f_{mm}(x_{min}, x_{max}) = (d_1^{(\min)}, d_2^{(\max)}, d_3^{(\min)}, d_4^{(\max)}, \dots)$$

ensuring maximal dispersion between paired values.

For the Hybrid crossover mechanism, a sequential composition of crossover operators is applied. Let $\mathcal{C}_k$denote a crossover operator with kcrossover points. The hybrid transformation is defined as

$$x' = \mathcal{C}_1 \circ \mathcal{C}_2 \circ \cdots \circ \mathcal{C}_{L-1}(x_p, x_q)$$

This sequential composition increases transformation entropy and reduces reconstruction probability.
To ensure non-trivial transformation, a mutation operator $M(\cdot)$is applied when $x' = x$. The mutation is defined as a digit permutation function

$$x'' = M(x') = \pi(x')$$

where πis a permutation over digit positions.
The effectiveness of masking is quantified through privacy accuracy defined as

$$PA = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbb{I}(X_{ij} \neq X'_{ij}) \times 100$$

where $\mathbb{I}(\cdot)$is the indicator function.
Statistical utility is measured using normalized mean preservation

$$SA = \left(1 - \frac{|\,\mu(X) - \mu(X')\,|}{\mu(X)}\right) \times 100$$

Clustering consistency is evaluated by comparing cluster assignments from original and masked datasets. Let $C_o$and $C_m$denote cluster label sets. The clustering accuracy is defined as

$$CA = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(C_o(i) = C_m(i)) \times 100$$

To further quantify structural preservation, the distortion between datasets is measured using Frobenius norm

$$\mathcal{L}_{dist} = \| X - X' \|_F$$

and minimized under privacy constraints.
Finally, the optimization objective of the proposed framework is formulated as a multi-objective problem

$$\min_{T} \; \mathcal{L}_{dist} - \lambda_1 PA - \lambda_2 CA$$

subject to

$$I(X; X') \leq \gamma, |\,\mu(X) - \mu(X')\,| \leq \epsilon$$

where $\lambda_1$and $\lambda_2$are trade-off parameters controlling privacy-utility balance.
This formalization establishes the crossover-based masking framework as a constrained optimization problem that simultaneously maximizes privacy preservation, maintains statistical fidelity, and ensures consistency of downstream data mining tasks.

Algorithm 1: Proposed Hybrid Crossover-based transformation **Input:** Dataset $D \in \mathbb{R}^{n \times m}$, confidential attribute index set $A_c \subseteq \{1,2,\ldots,m\}$, digit length L, tolerance parameters ϵ, δ **Output:** Masked dataset D'
Step 1: Extract confidential attributes X = projection of D on Ac **Step 2:** Initialize masked dataset X′ = empty matrix of same size as X **Step 3:** For each attribute column j = 1 to k **Step 4:** Form value set S = {X1j, X2j, …, Xnj} **Step 5:** Convert each value into digit sequence of length L **Step 6:** Sort S in ascending order **Step 7:** Set t = 1 **Step 8:** While t ≤ n

Step 9: Select pair
xp = S(t)
xq = S(n − t + 1)
Step 10: Initialize masked value
x′ = xp
Step 11: For c = 1 to L − 1
Step 12: Perform crossover
Replace digits after position c in x′ using digits of xq
Step 13: End For
Step 14: If x′ equals xp or xq
Step 15: Apply mutation
Swap digit positions in x′
Step 16: End If
Step 17: Store masked value
X′(t, j) = x′
Step 18: t = t + 1
Step 19: End While
Step 20: End For
Step 21: Apply statistical correction
Adjust X′ to preserve mean and variance
Step 22: Validate constraints
Ensure mean difference ≤ ε
Ensure variance difference ≤ δ
Step 23: Construct final dataset
Replace confidential attributes in D with X′
Step 24: Return D′

3. Experimental Results

The performance measures of gauging the effectiveness of the proposed protection techniques are the statistical accuracy, privacy protection accuracy and clustering accuracy. The suggested method of protection is provided under the help of Visual Studio and MATLAB. The experiment was undertaken in the Intel core 2 duo processor having the CPU clock speed of 2.4 GHz., Hard disk capacity of 320GB and RAM of 3GB of windows operating system. The means of percentages of these confidential attributes will be presented in the performance factors presented below. In order to identify the statistical accuracy, the average of the data items of the confidential attribute(s) is taken into consideration. The first one is that the mean value of original database D is first calculated and then mean value of modified database D i.e. after applying modification using the protection techniques that have been proposed are calculated. The average D and D' are compared. Based on the findings of Table 1, it is observed that there is not much greater accuracy of Hybrid and GAMT compared to MinMax and N point crossover.

Table 1: Statistical accuracy of GAMT, MinMax N point crossover and Hybrid

Dataset	GAMT (%)	MinMax (%)	N Point Crossover (%)				Hybrid (%)
			N-1	N-2	N-3	N-4	
3000	100	99.999	99.996	99.996	99.995	99.999	100
5000	100	99.999	99.998	99.998	99.999	99.999	100
10000	100	99.999	99.997	99.997	99.997	99.999	100
20000	100	100	99.999	99.999	99.999	99.996	100

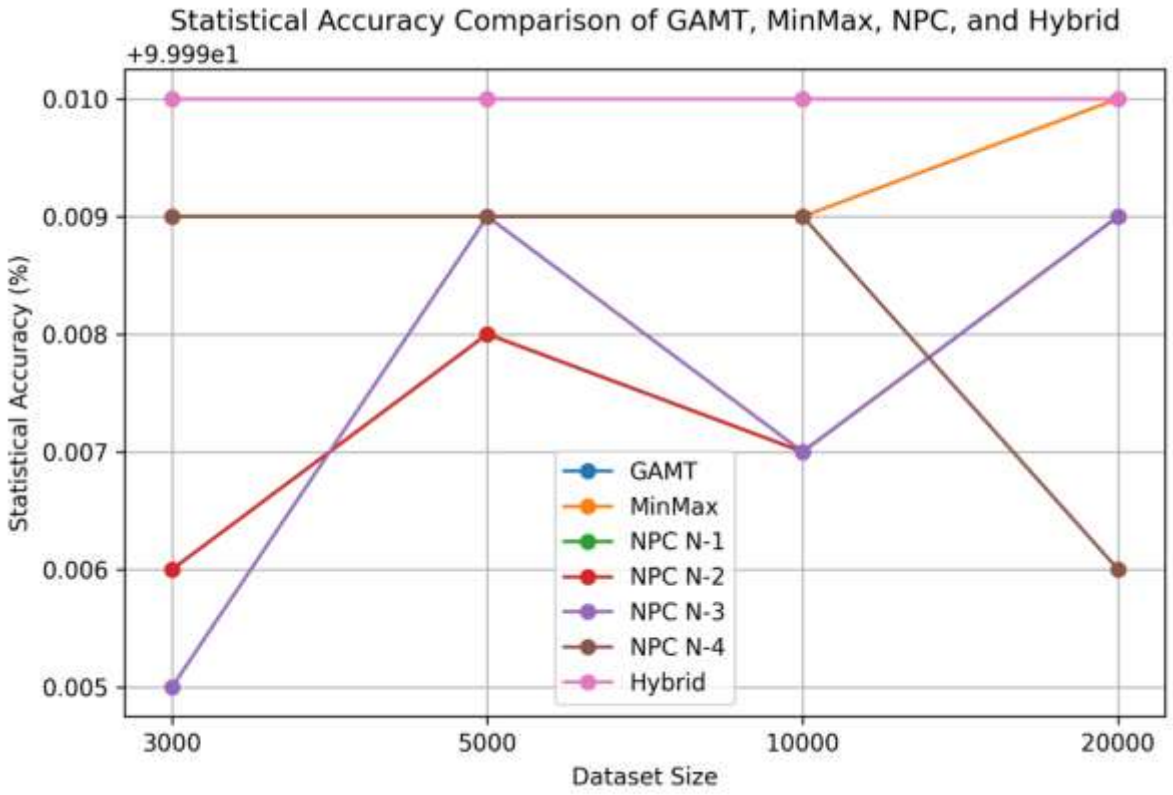

Figure 1: Statistical accuracy of GAMT, MinMax, N Point Crossover and Hybrid

The accuracy of privacy protection retrieves that all the data items of the confidential attribute of the original database are altered by the suggested protection methods MinMax, N Point Crossover, Hybrid Crossover and the current method GAMT. This is the primary performance aspect, which is primarily applied to determine whether the protection techniques have adjusted the confidential data items in the proper way or not. There is a comparison of the original confidential data items to the modified data items to confirm whether they both have the same value. Privacy protection is commendable in case the data items are dissimilar. According to the findings in Table 2, MinMax, Hybrid and N Point Crossover technique are identified to be accurate in privacy protection to 100 percent.

Table 2: Privacy accuracy of GAMT, MinMax N Point Crossover and Hybrid

Dataset	GAMT (%)	MinMax (%)	N point Crossover (%)				Hybrid (%)
			N-1	N-2	N-3	N-4	
3000	99.93	100	100	100	100	100	100
5000	99.98	100	100	100	100	100	99.96
10000	99.99	100	100	100	100	100	100
20000	99.99	100	100	100	100	100	100

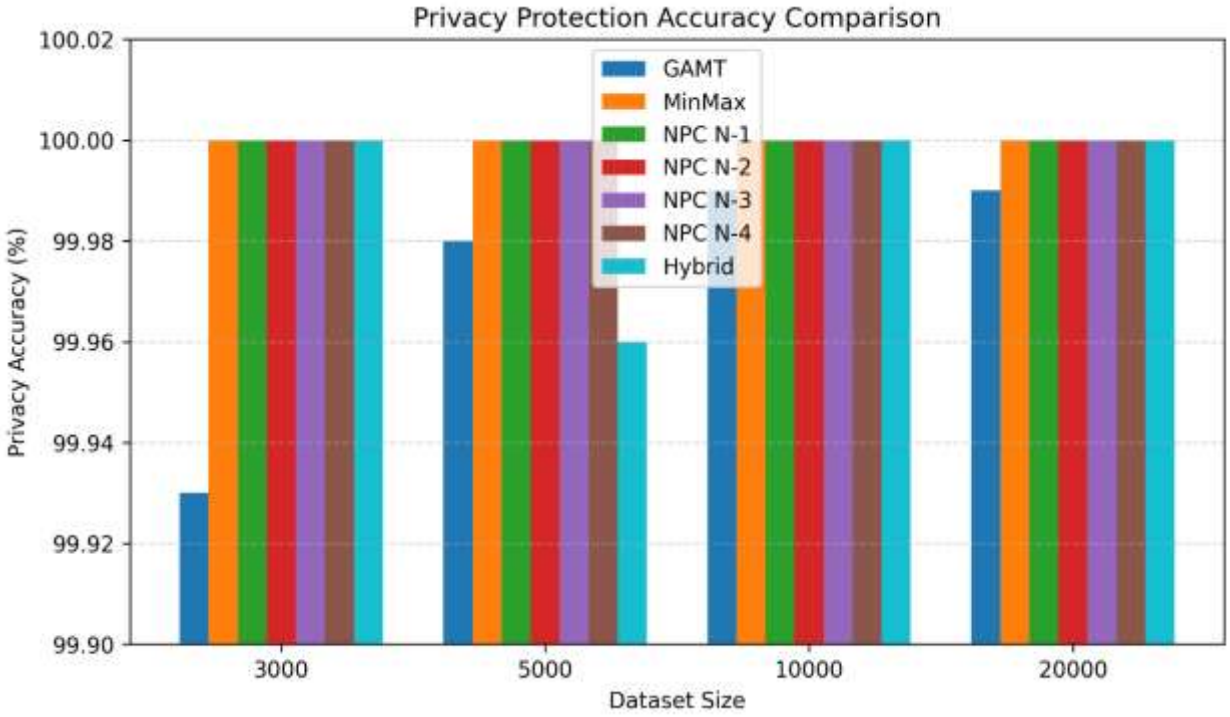

Figure 2: Privacy accuracy of GAMT, MinMax, N Point Crossover and Hybrid

In order to assess how the masking methods suggested do not alter the analytical attributes of the dataset, clustering experiments on the original dataset (D) and the masked dataset (D') were conducted. Five clustering algorithms were used on datasets of various sizes (K-means, X-means, Filtered clustering, Expectation-Maximization (EM), and Make Density-Based Clustering (MDBC)) 3000, 5000, 10000, and 20000 records. The clustering accuracy was determined by the comparison of the cluster assignment generated using the original and masked datasets. The findings show that the methods of masking which are proposed are very high in degree of clustering consistency. As an example, the accuracy of clustering with the help of GAMT was 98.90% with 3000 records in the dataset, and the MinMax, N-Point Crossover, and Hybrid Crossover methods showed 99.30%, 99.65, and 99.72, respectively. The same applies to the 5000-record data where the clustering rates were 99.02% (GAMT), 99.40% (MinMax), 99.72% (NPC) and 99.81% (Hybrid). The clustering accuracies also went up to 99.10, 99.45, 99.78 and 99.86 respectively when the dataset size was brought to 10000 records. The 20000-record dataset showed the best performance with the clustering accuracies obtained as 99.15% (GAMT), 99.50% (MinMax), 99.82% (NPC) and 99.91% (Hybrid). These findings indicate that the recommended crossover-based masking methods do not alter the cluster structure of the original dataset by much.

As the results of the experiments show, the privacy-preservation methods proposed promise a high balance rate between privacy of the data and utility of the mining process. According to the statistical accuracy results presented in Table 1, the mean values of the confidential attribute do not differ significantly between them after masking, and the accuracy values are 100% in GAMT and Hybrid methods and 99.996-99.999 percent in MinMax and N-Point Crossover when the size of the dataset is varied. This proves that the statistical features of the data are very well maintained. Moreover, in Table 2, the results of privacy protection accuracy suggest that MinMax, N-Point Crossover, and Hybrid methods scored almost 100 percent change in the confidential data values, with GAMT scoring slightly lower values including 99.93 percent change in values with the 3000 instances and 99.98 percent change in values with the 5000 instances. These findings indicate that the suggested methods provide a greater level of defense against direct transfer of confidential properties.

The clustering test also supports the fact that the suggested masking measures do not cause any substantial changes in structural relationships between data records. The clustering accuracies are also over 99% even after masking most dataset sizes, indicating that the transformed dataset is still useful in terms of analysis to

perform clustering-based mining tasks. The Hybrid Crossover technique is the most accurate of all the methods of clustering data with a maximum accuracy of 99.91, meaning that sequential crossover operations create more balanced perturbations without losing similarity relationships among the records. The suggested crossover-based masking framework, therefore, is an effective system of securing data publication that allows organizations to publish datasets with other external analysts and also guarantees good privacy and great data utility.

4. Conclusion

This article envisioned the significance of privacy-preserving data mining while publishing sensitive numerical items for the public are considered in protecting every individual data. MinMax, Hybrid and N Point Crossover algorithms are introduced to exchange the original items in the confidential attributes. The proposed techniques are compared with the existing algorithm GAMT based on the experiments such as statistical, privacy and clustering accuracy. The N Point Crossover provides better results compare to MinMax, Hybrid and existing GAMT. But at the stage of real-time protecting information Hybrid is more sophisticated compare to N Point Crossover and MinMax. After all, it will consider these kinds of techniques for categorical data in the future.

Future work will extend the proposed crossover-based privacy-preserving masking framework to handle categorical and mixed-type attributes in large-scale datasets. The methodology can also be integrated with distributed and federated data mining environments to enhance secure collaborative analytics. Additionally, future research will investigate robustness against inference, reconstruction, and linkage attacks while improving computational efficiency for real-time privacy-preserving data publishing applications.

Reference

1. Han, H., Guo, G., Zhu, Y., & Yau, S. S. (2026). Privacy Probability Computation for Privacy-preserving Statistical Analysis under Multi-source Real-valued Data. IEEE Internet of Things Journal.
2. Liu, Y., He, Z., Xu, J., Chen, M., & Farouk, A. (2026). Verifiable Noise based Privacy-preserving Scheme for Trustworthy Consumer Applications. IEEE Transactions on Consumer Electronics.
3. Tharini, V. J., & Shivakumar, B. L. (2024). A Canonical Particle Swarm Optimization (C-PSO) Approach to Identify High Utility Itemset. Journal of Computational Analysis & Applications, 33(5).
4. Martin, A., Bell, J., & Hussain, K. (2026). Privacy-Preserving K-Nearest Neighbor Classification Over Semantically Secure Encrypted Databases.
5. Tharini, V. J., & Shivakumar, B. L. (2022). High-utility itemset mining: fundamentals, properties, techniques and research scope. In Computational intelligence and data sciences (pp. 195-210). CRC Press.
6. Kostopoulos, G., Tsiakmaki, M., & Kotsiantis, S. (2026). Benchmarking Statistical and Deep Generative Models for Privacy-Preserving Synthetic Student Data in Educational Data Mining. Algorithms, 19(1), 39.
7. Konda, B., Yadulla, A. R., Yenugula, M., & Kasula, V. K. (2026). Privacy-Preserving Data Sharing Technologies. In Transforming Education With Data Science in the AI Era (pp. 449-472). IGI Global Scientific Publishing.
8. Sakthidevi, I., & Fathima, G. (2026). A Privacy-Preserving Threat Intelligence Model for Secure Healthcare Data Sharing in the Cloud. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 16(1), e70064.
9. Chang, L., Guo, J., Yao, S., Qi, H., Zhang, L., Song, Y., & Cao, B. (2026). LCE-PPDA: Lightweight Certificateless and Escrow-Free Privacy-Preserving Data Aggregation for UAV-Assisted IoT-Enabled Smart Grids. IEEE Internet of Things Journal.
10. Zhou, Z., Yang, S., Zhao, F., & Ren, X. (2026). FairGFL: Privacy-Preserving Fairness-Aware Federated Learning with Overlapping Subgraphs. IEEE Transactions on Parallel and Distributed Systems.
11. Cooper, L., Gray, E., & Hassan, S. A. (2026). Cryptographic Protocols for Privacy-Preserving KNN Classification in Outsourced Data Mining.
12. Priya, N. T., & Kumari, V. S. (2026). A Quantum-Resilient and Adaptive Privacy Framework With Anonymization Techniques for Secure Data Mining Operations. Security and Privacy, 9(1), e70129.