Enhancing Privacy Protection for Sensitive Data Using Differential Privacy and Generative Adversarial Networks

S. Vidivelli[1], Leelankitha Kanchanapalli[2,] R. Manikandan[3,*], S. Magesh[4]
[1,2,3]School of Computing, SASTRA Deemed to be University, Thanjavur, Tamilnadu, India
[4]Department of Computer Science and Engineering, Dr. M. G. R Educational and Research Institute, Chennai, India
Email id: vidieng@gmail.com, 125018041@sastra.ac.in, srmanimt75@gmail.com, techiemagesh@gmail.com

Abstract

Efficient anonymization techniques have become one of the most critical research issues for the data privacy community. A large number of datasets are distributed and shared on different platforms to be utilized by communities across various industries. Unfortunately, the traditional methods of Anonymizing Data fall short of maintaining data privacy, which led to the exploration option for Synthetic Data Generation. Deep learning has been widely accepted for its high-accuracy privacy-preserving mechanism. In light of this, researchers have relied on synthetic data generated by deep learning to protect the original data from being leaked and avoid resulting attacks. Relationships between many features Deep learning models can effectively capture relationships between many features from multiple tables. The study introduces a new method based on using Differential privacy (DP) with Generative Adversarial Networks (GANs). Quasi Identifiers are recognized for the adult dataset and then separated from Sensitive Attributes, subsequently applying differential privacy anonymization mechanisms over them by adding two different noise methods, Laplacian noise and Gaussian Noise, to obtain encoded data. These anonymized records can be used as actual training data in the GANS Model to evaluate both Laplacian mechanism performance and Gaussian mechanism performance. The proposed method achieved an accuracy of 85.2%, demonstrating its efficiency in maintaining data utility while enhancing data privacy.

Keywords

Data Utility. Differential Privacy. Anonymization. Generative Adversarial Networks (GAN). Synthetic data generation. Quasi identifiers (QI). Sensitive attributes (SAs). Laplacian Noise. Gaussian Noise. Deep Learning. Data Anonymization

Abbreviations

- QI: Quasi-Identifier
- SA: Sensitive Attribute
- DP = Differential Privacy
- g: Generator model
- d: Discriminator model
- m: Noise prior
- p: Generator distribution

- r: Target distribution
- X, Y: Distribution function

Greek Letters

- ε (epsilon): Privacy measure
- δ (delta): Failure rate of measure
- θ (theta): Sensitivity of the data
- σ (sigma): Standard deviation

1 Introduction

When processing sensitive data from the adult dataset, we are dealing with highly personal information that uniquely identifies a person.[1] We utilize this dataset as a standard socioeconomic label and reference for a variety of investigations since it contains thorough information on fundamental employment records, sociodemographic variables, and educational background. Privacy Matters: The best adequate protection of personal information is essential for ensuring a privacy-secure society. Not only must the data points in the dataset be as complete as possible, but we also need this information to be handled securely. Adult data is primarily available to researchers and policymakers, facilitating access to socioeconomic analysis for informed decision-making. The danger of information leakage in datasets, including personal and sensitive information, is considerable, and the consequences can be severe. Sensitive data may be disclosed to unauthorized organizations, resulting in the misuse of personal confidential information and substantial financial damages. Information, once disclosed, cannot be withdrawn or kept secret, hence the name inside information. Identity disclosure refers to the association of an individual with a record in a dataset.

On the other hand, attribute disclosure occurs when there is an affiliation of an individual with a record, but a previously unknown attribute about him/her is described. Such solutions are mainly intended to resolve these types of disclosures. However, the typical approaches of data anonymization [2] seen in the literature typically sacrifice data usability for privacy. In this framework, Laplacian noise and Gaussian noise are two standard methods used for privacy enhancement. Thus, with the help of synthetic data generation, it became possible to find a practical approach to making the data useful while preserving their confidentiality. GANs are basically a form of neural network that can generate synthetic data: generate artificial datasets that preserve sample structure of original dataset.

This capability makes it possible for GANs to identify patterns in data, and this makes data generated by GANs suitable for various analyses. Compared with the previous generation methods, GANs provided significantly superior results for data anonymization compared to variational autoencoders, restricted Boltzmann machines, and traditional k-anonymity processes. For instance, differential privacy uses controlled noise to shield individual data, while federated learning allows models to use multiple decentralized data sources yet does not share the data. In SMP, computations are done on encrypted data, which provides much protection in terms of privacy. In the case of privacy-preserving data generation, the main advantages of GANs include: (i) They preserve privacy by mitigating the risks associated with classical anonymization. (ii) They improve data utility. Unlike traditional

anonymization approaches, which allow attackers to extract information from anonymized data, synthetic data generated by GANs makes it harder for attackers to infer sensitive information.

The paper describes **Enhancing Privacy Protection for Sensitive Data Using Differential Privacy and Generative Adversarial Networks**, a novel method for prioritizing the protection of attributes that cause privacy leaks in the adult dataset.[3, 4] According to the rules of the QI approach, the QI and sensitive attributes in the adult dataset are classified. Differential privacy is applied to enhance the data using QI and sensitive attributes, and then the anonymized data is used to train the GAN model. The performance of the proposed Fortified-GAN method was further tested using the benchmark dataset, namely the adult dataset, which was obtained from the UCI Machine learning resource. The format of the paper is as follows: Section 2 presents a brief literature review of privacy-securing approaches and synthetic data generation. Section 3 describes the method that has been applied here, primarily the separation of QI and SA characteristics and the introduction of noise. In section 4, the authors include their results concerning the application of Laplacian and Gaussian noise, together with the research done in applying them for training GANs. Finally, Section 5 offers the article with the conclusion and an outline of the study's prospects.

Key Contributions:

The contribution of the study is as follows: (i) Introduces a novel approach of privacy-preserving technique combining differential privacy with GANs. (ii) Evaluates the performance of Laplacian vs. Gaussian noise in synthetic data generation. (iii) Demonstrates superior data utility preservation with Laplacian noise. (iv) Proposes a comprehensive evaluation framework using precision, recall, F1 score, and accuracy.

2 Related Work

In the field of sensitive data privacy, there are such objectives as confidentiality and integrity of the information. In sensitive data protection, the confidentiality and integrity of the data while maintaining the value of data are paramount in different fields. Some standard methods that academics have discussed concerning the balancing of privacy risks that result from data sharing include differential privacy mechanisms. Deletion and dissociation activities have been applied to banking records, demography data, and telecommunications databases, indicating the importance of the conception of anonymization and de-identification concerning the contradiction of data utility and privacy [5,6]. The advancements of machine learning through GANs generate synthetic data, which retain the statistical properties but keep the individual information private; recent attempts have been made to address the privacy issues and develop secure data-sharing techniques.

2.1 Privacy-preserving techniques

In the context of our study on privacy enhancement using GANs and differential privacy techniques, one of the classical anonymization algorithms was designed to protect identification in adult data by removing direct identifiers. There are many more techniques for anonymization of the data. However, these techniques are vulnerable to attacks such as background information, similarity, skewness, and probabilistic inference attacks. While these solutions prioritize data privacy, they usually compromise data utility. Differential privacy, developed by Dwork et

al.,[7,8,9,10]. addresses this trade-off by enhancing data utility while retaining privacy. It enables the interchange of meaningful statistical information from sensitive datasets, whereas older methods typically result in lower data accuracy. Several methods for anonymization have been proposed as part of our study on privacy enhancement using GANs with differential privacy strategies. HIDE, for example, maximizes data utility by employing a conditional random field-based method for structured data extraction and k-anonymization. Loukides et al,.[11]employed disassociation to separate records and prevent data linkage. To address anonymization difficulties while keeping data utility, synthetic data generation using deep learning models has grown in favour. Acs et al.,[12] used GAN to cluster and generate factual synthetic data despite the challenges of retaining resistance against medical data attacks. Kaushik et al. designed VGAN (Variance GAN) method to decrease the disparity between the model and actual data variability; this technique was influential in the medical expenditure prediction but revealed poor robustness to network attack. Furthermore, Li et al,.[13]studied feature vector GAN techniques for targeting learning-based classifiers, demonstrating performance disparities between attack types. Hitaj et al. [14] studied information leakage possibilities in collaborative deep learning and show that GANs can use shared model updates in order to reconstruct sensitive training data and, consequently, leak privacy in deep learning models. Fredrikson et al.[15]provided a case study about the privacy risks in pharmacogenetics more precisely in personalized warfarin dosing, where adversarial access to models results in leakage of sensitive genetic data. Manning and Haglin [16] presented a new algorithm in 2005 that improves statistical disclosure analysis through the detection of minimal sample uniques, which has been an important factor in the access of data disclosure risk in datasets. Arjovsky and Bottou (2017) [17] shed light on the aspect of GAN training by underlining that principled methods are needed to highlight stability and convergence issues plaguing the training process of the generative model. Zhang et al. [18] proposed PrivBayes, a framework using Bayesian networks to release data with privacy, providing a high level of utility; these methods ensure data privacy by using differential privacy methods, which prevent disclosure risks.

3 Proposed Work

We provide privacy preserving GAN, a unique way to increase privacy and data value in sensitive datasets by combining Generative Adversarial Networks (GANs) and differential privacy strategies. As the number of sensitive data increases, maintaining privacy becomes more difficult due to deep data correlations. To overcome this issue, we will conduct comprehensive data analysis using a deep neural network model. This will operate in three major phases: (i) Identify Quasi-Identifiers (QI) and classify remaining attributes as Sensitive Attributes (SA); (ii) Apply dp techniques, specifically f-differential privacy, to QI attributes while integrating them with SA attributes to preserve data utility; and (iii) Use differential privacy techniques to obtain synthetic data with improved utility. By detecting QI traits in the adult dataset and using noise approaches like Laplacian and Gaussian noise, the strategy ensures privacy while keeping data utility. These anonymized QI features are then used to train GAN in order to generate synthetic data that will assist in mitigating this privacy threat or, re-identification or other attacks.

Algorithm of Proposed Model:

Input: Dataset D with attributes, Privacy parameters ε (epsilon), δ (delta).
Output: Synthetic dataset S

1. Function IdentifyAttributes(D):
   a. QI ← IdentifyQuasiIdentifiers(D)
   b. SA ← IdentifySensitiveAttributes(D)
   c. Return QI, SA
2. Function ApplyDifferentialPrivacy(QI, SA, ε, δ):
   a. A ← Initialize Anonymized Dataset
   b. For each attribute q in QI:
      i. Δq ← CalculateSensitivity(q)
      ii. Add Laplacian noise:
         - q' ← q + Laplace(0, Δq/ε)
      iii. Add Gaussian noise:
         - σ ← Δq/ε
         - q" ← q + N(0, σ^2)
      iv. Append q' and q" to A
   c. Combine A with unaltered SA to form A'
   d. Return A'
3. Function TrainGAN(A'):
   a. Initialize generator g and discriminator d
   b. For each epoch:
      i. For each batch of real data R from A':
         - Generate synthetic data G from g
         - Train d on R and G
         - Update d based on loss
      ii. Update g based on discriminator feedback
   c. Return synthetic dataset S from g
4. Main:
   a. D ← LoadDataset()
   b. QI, SA ← Identify Attributes(D)
   c. A' ← ApplyDifferentialPrivacy(QI, SA, ε, δ)
   d. S ← TrainGAN(A')
   e. Return S

3.1 QI and SA attributes

The model's first process is to detect QI attributes in the adult dataset. Adult datasets sometimes provide third-party direct identifiers. These types are usually not used when disseminating data to other people or when conducting research work**. Fig.1** describes the work flow of the proposed model.

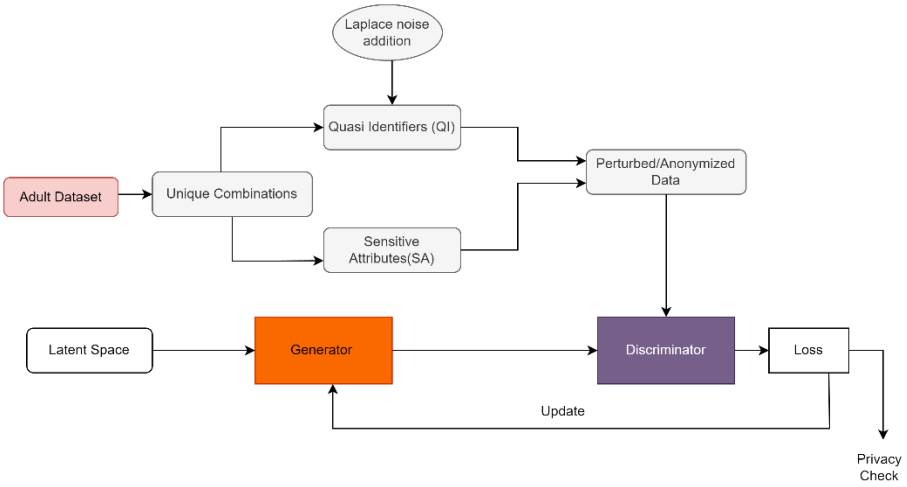


Fig. 1 Overview of proposed model

Following the recalculations aimed at removing direct identifiers, remaining attributes are called the sensitive attributes (SA) and the quasi-identifiers (QI). SA traits refer to information that people may want to conceal. At the same time, QI attributes are features derived from the union of dimensions aggregated into unique identifiers that can be used to re-identify people given to other databases. When it comes to privacy issues associated with QI qualities, anonymization techniques are applied only to certain aspects, and this is where the focus is only on the identified QI qualities; SA qualities remain as is. The specifics of this strategy comprise the rationale behind the effective augmentation of the data value in the context of the adult dataset without undermining these subjects' privacy based on the research. This implies that the anonymization technique only takes place to indict QI attributes that experienced a breach of privacy. This achieves the optimal use of limited data while at same time ensuring people's privacy is upheld. Observing the attributes of the adult dataset, we identified that it contained quasi-identifiers (QI) [19] and sensitive attributes (SA) by analyzing different combinations of attributes. SPOA characteristics are the attributes that, when added to other datasets, can help trace the subject. These include the age, gender, race, and ZIP code into which a person or recipient falls into. At the same time, SA contains information that can be assessed as personal, including income, education level, and occupation. Knowledge of these characteristics is essential in the selection of proper anonymization techniques that will enhance the data's usefulness while still respecting the right to privacy. It also eliminates any ambiguity on overloading the QI with tasks that will ensure our technique reduces risk of re-identification without letting go of SA data, which is valuable in future studies. **Table 1** presents the metadata for the dataset, showcasing the first six rows of the 15 attributes. **Table 2** and **Table 3** detail the quasi-identifier (QI) and sensitive attributes (SA) identified, respectively.

Age	workclass	fnlwgt	education	edu-num	Marital Status	Occupation
39	state-gov	77516	Bachelors	13	Never-married	Adm- clerical
50	self-emp	83311	bachelors	13	Married-civ-spouse	Exec-managerial
38	private	215646	HS-grad	9	Divorced	Handlers-Cleaners
53	private	234721	11th	7	Married-civ-spouse	Handlers- Cleaners
28	private	338409	Bachelors	13	Married-civ-spouse	Prof-special
37	private	284582	Masters	14	Married-civ-spouse	Exec-managerial

relationship	race	sex	capital-gain	capital-loss	hours per-week	native country	income
Not family	W	M	2174	0	40	US	<=50K
Husband	W	M	0	0	13	US	<=50K
Not family	W	M	0	0	40	US	<=50K
Husband	B	M	0	0	40	US	<=50K
Wife	B	F	0	0	40	Cuba	<=50K
Wife	W	F	0	0	40	US	<=50K

Table 1: Meta Data with 15 attributes(First 6 rows)

Age	capital-gain	fnlwgt	capital-loss	hours per-week	native country
39	2174	77516	0	40	US
50	0	83311	0	13	US
38	0	215646	0	40	US
53	0	234721	0	40	,US
28	0	338409	0	40	Cuba
37	0	284582	0	40	US

Table 2: Quasi Identifiers(QI)

work class	education	edu-num	Marital Status	Occupation	relationship	race	sex	income
state-gov	Bachelors	13	Never-married	Adm clerical	Not family	W	M	<=50K
self-emp	bachelors	13	Married-civ-spouse	Exec-managerial	Husband	W	M	<=50K
private	HS-gard	9	Divorced	Handlers-Cleaners	Not family	W	M	<=50K
private	11th	7	Married-civ-spouse	Handlers-Cleaners	Husband	B	M	<=50K
private	Bachelors	13	Married-civ-spouse	Prof-special	Wife	B	F	<=50K
private	Masters	14	Married-civ-spouse	Exec-managerial	Wife	W	F	<=50K

Table 3: Sensitive Attributes(SA)

3.2 Noise Addition

The introduction of noise at an early stage of the model ensures that during the training and testing of the model, the adversaries cannot easily extract individual information. Compassionate data in the sample case, such as adult data, can also be appropriately dealt with so that data utility is not compromised. Differential privacy, as defined above, is one of the techniques used in data anonymization. The differential privacy (DP) guarantees privacy because it prevents attackers from outputting precise information regarding individuals from the dataset.

3.2.1 Differential Privacy and Noise Mechanisms

A randomized procedure. X is ϵ-differentially private if

$$Pr[X(y) \in T] \leq e\hat{} \epsilon \ Pr[X(y') \in E] \quad (1)$$

for all pairs of adjacent databases b and b′ (which differ in only one individual record) and for all sets E ⊆ Supp(X(b)) ∪ Supp(X(b′)) where ϵ is the privacy measure, as evaluated as in Eq. (1).

(ϵ,δ)-Differential Privacy

A randomized method X provides (ϵ,δ)-differential privacy if for all pairings of adjacent databases b and b′ and all E ⊆ Range(X)

$$Pr[X(b) \in E] \leq e\hat{}\epsilon \ Pr[X(b') \in E] + \delta \quad (2)$$

Here, as evaluated as in Eq.(2), δ is the failure rate for which the privacy value provided by ϵ does not apply.

Laplace Mechanism

Laplacian noise can be added to a statistic θ using the Laplace mechanism. The formula to add Laplacian noise is: $\theta' = \theta + Lap(\{\Delta\theta\}/\{\epsilon\})$ (3)

Where θ is the original statistic, θ′ is the noisy statistic, as evaluated as in Eq.(3) Lap denotes the Laplace distribution, Δθ is the sensitivity, and ϵ is the privacy measure.

Gaussian Mechanism

Gaussian noise can be added to a statistic θ via the Gaussian mechanism. The formula to add Gaussian noise is: $\theta' = \theta + N(0, \sigma^2)$ (4)

Where θ is the original statistic, θ′ is the noisy statistic, , as evaluated as in Eq.(4) N(0, σ^2) denotes the Gaussian distribution.

The noise introduced here is from the above Gaussian distribution f-DP [20], the short form of which is GDP: Gaussian Differential Privacy. It presented a method that estimates the half width at half maximum, given the result of testing of 2 shifted Gaussian distributions. The method guarantees that identifying any of the two contiguous datasets is as hard as distinguishing two Gaussian distributions based on one grain. It can further be defined that if in two neighbouring datasets b and b′ and distributions P and Q, , as evaluated as in Eq.(5) the trade-off function U satisfies

$$E\big(X(b), X(b')\big) \geq U \quad (5)$$

This definition of Gaussian DP (μ-GDP) uses the mean of a unit-variance to give an effective privacy guarantee. In our research, we apply these ideas to improve the privacy of the adult dataset by detecting and anonymizing Quasi-Identifiers (QIs). Hence, through the employment of Gaussian noise, we address privacy issues and make them robust with regard to QI characteristics. The Gaussian noise mechanism possesses standard deviation σ, as evaluated as in Eq.(6)

$$\sigma^2 = sens(\theta)^2/\mu^2 \quad (6)$$

ensures that the result is private and still useful in processing the data. This is included in our model and allows us to create synthetic data to minimize the privacy problem. Thus, the idea of adding Gaussian noise, not into the whole dataset but into the QIs, is closer to providing an appropriate level of data protection while maintaining usefulness. When deriving the synthetic data using the above GAN, useful QI attributes are augmented with sensitive attributes in an anonymous format. **Fig.2** compares the distributions of real data with data perturbed by Gaussian and Laplacian noise. **Fig.3** shows the Mean Squared Error (MSE) for both noise types, highlighting

their impact on data accuracy.

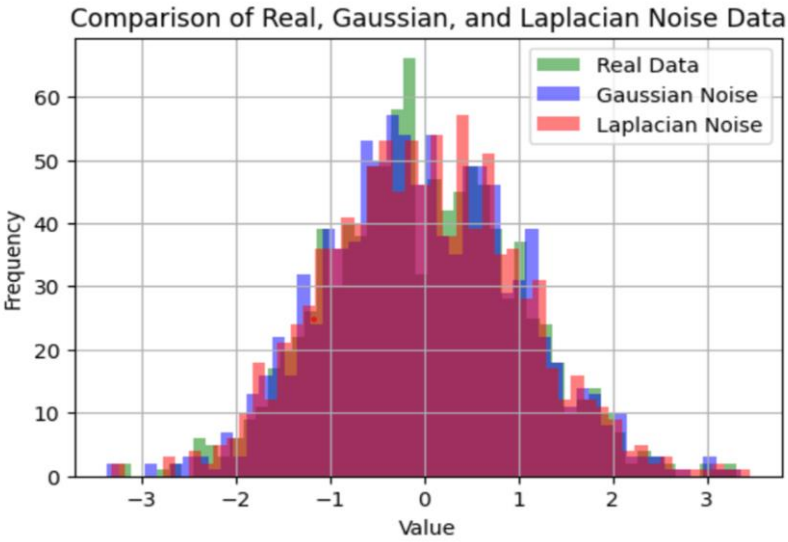


Fig.2: Comparison of accurate Gaussian and Laplacian noise data.

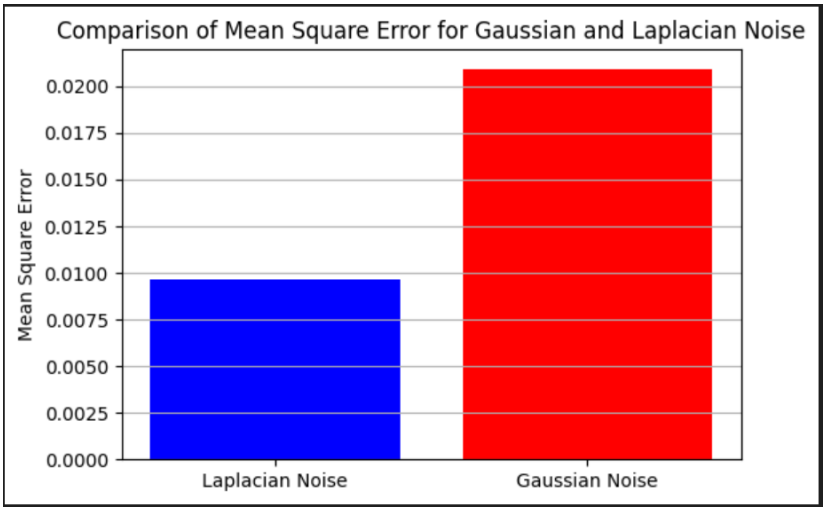


Fig.3: MSE for both Gaussian and Laplacian noise data.

3.3 Enhancing Protection for Sensitive Data Using GANs

In the project, sensitive attributes (SA) and quasi-identifier (QI) values are used to anonymize medical data inputs before being fed into the Generative Adversarial Network (GAN).[21] GANs are used in cybersecurity to guard against adversarial assaults, as well as to protect privacy by anonymizing sensitive data.

- **Generator (g)**: Gets synthetic data that attempts to mimic the actual data distribution and tricks the discriminator into identifying it as accurate.
- **Discriminator (d)**: Evaluates samples by discriminating between genuine data using a training set and data obtained by the generator.

The generator (g) learns to produce the data instances that closely look like distribution of original data, and the discriminator (d) learns the difference between actual and generated data. This process is regulated by a joint probability p(r,s), where r represents the actual data instances and s means the synthetic data produced by g. The discriminator seeks to maximize its capacity to differentiate between r and s, whereas the generator seeks to minimize such distinguishability. GANs optimize both the generator and discriminator using the same loss function: $p_{data}(r)$ for real data and pgen(S) for produced data.

Discriminator Loss (C(D)): Maximises the discriminator's capacity to accurately identify real and created data instances. The discriminator reduces it as evaluated as in Eq.(7)

- $$C(d) = -Y_{r\sim p_{data}(r)}[log\ d(r)] - -Y_{s\sim p_{gen}(s)}[log(1 - d(s)] \quad (7)$$

Generator Loss (C(G)): Reduces the distinguishability of produced data from the real data. The generator maximizes it as evaluated as in Eq.(8)

- $$C(G) = Y_{s\sim p_{gen}(s)}[log(1 - d(s))] \quad (8)$$

The GAN framework receives anonymized adult data, which includes QI and SA properties. The project's goal is to create synthetic data that preserves privacy while preserving utility by incorporating differential privacy approaches into the QI characteristics and using KL divergence as a divergence measure.

3.3.1 Divergence measures

Many techniques have been developed to reduce the risk of data re-identification and disclosure of attributes in privacy-preserving data analysis. Such technique is k-

anonymity,[22] guarantees that no record in any dataset can be distinguished from at least. The outcome is k−1 other records based on quasi-identifiers, which provide the base level of anonymity. However, k-anonymity alone cannot protect against higher-level privacy attacks; for example, attribute disclosure, where adversaries can make the most out of homogeneity within equivalence classes. ℓ-Diversity[23] enhances this requirement through the extension of k-anonymity, as it enforces each equivalence class to possess a diverse set of values for sensitive attributes, thus providing stronger protection against any kind of attribute inference. Despite all these developments, classical anonymization techniques remain highly susceptible to those sets of new and emerging privacy threats known as membership inference[24] and model inversion attacks[25]. Unlike traditional methods, our model addresses a novel approach that integrates differential privacy [26] with the Generative Adversarial Networks to enhance privacy protection while retaining good utility in data.

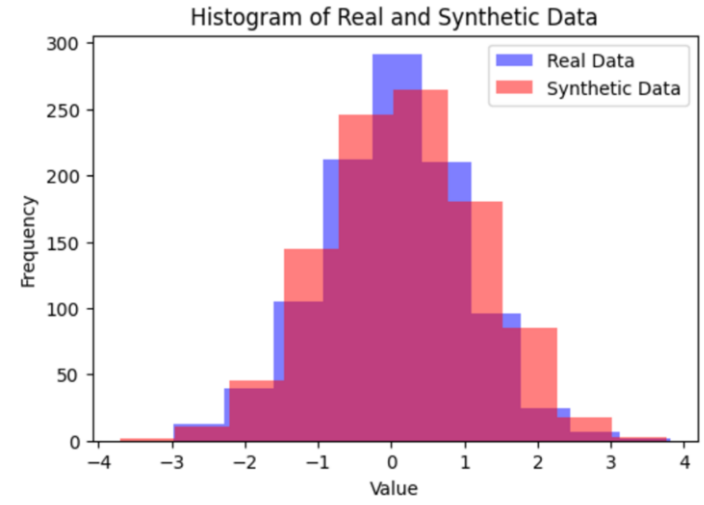
KL divergence is preferred in our research over other divergence measures in the project due to its precise optimization capabilities and suitability for high-dimensional data distributions like those found in the adult dataset. KL divergence effectively captures the difference between the actual and generated data distributions, allowing for fast gradient-based updates to the GAN [27]. Furthermore, KL divergence's sensitivity to variations in probability distributions enables the GAN to focus on lowering significant discrepancies, improving the privacy and utility balance required for anonymizing quasi-identifiers while keeping the integrity of critical attributes. The objective of the generator g is the same as the distribution p of the actual data, which it is trying to mimic. At the same time, the discriminator d is trained to detect the real data from the generated data. Using the above calculus, KL divergence is employed to measure the divergence of the generated distribution q from the actual distribution p.

The generator and discriminator loss functions with KL divergence are defined as follows: d(x) is the probability that x belongs to the actual data distribution p, and g(z) is the data generated from the latent z as evaluated as in Eq.(9) and Eq.(10)

$$L_g = d_{KL}(p \parallel q) = E_{x\sim p)}\left[log\left(\frac{\{p(x)\}}{\{q(x)\}}\right)\right] \quad (9)$$

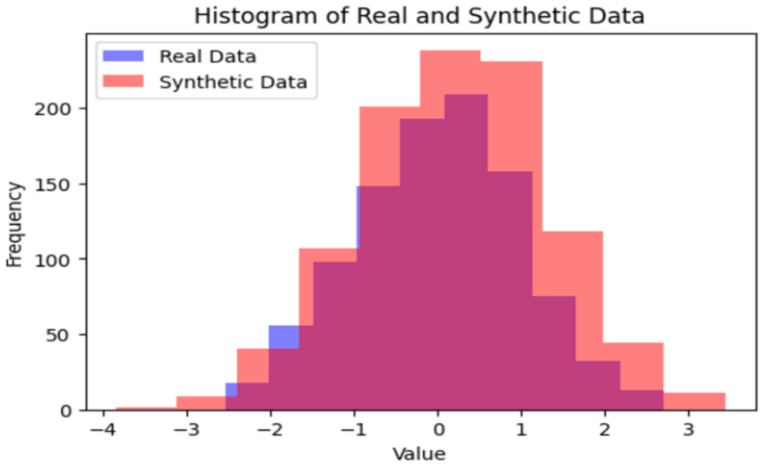
$$L_d = d_{KL}(p \parallel q) = E_{x\sim p)}\left[log\ D(x)\right] - E_{z\sim q)}\ log\ (1 - d(g(z)))\right] \quad (10)$$

The system uses KL divergence to assess the divergence between real and generated data distribution effectively. This strategy overcomes the drawbacks of current divergence metrics by introducing a solid mechanism for balancing privacy and data value. **Fig.4** and **Fig.5** demonstrate the utility of the real data compared to the synthetic data generated by SDG-GAN using Laplacian and Gaussian noise.



KL Divergence: 0.3579897524058031

Fig.4: Utility for Real and SDG-GAN(laplace)



KL Divergence: 0.9181910075205539

Fig.5: Utility for Real and SDG-GAN(Gaussian)

4 Implementation

The adult dataset is a mix of numerical and categorical variables. The GAN uses num. In this way, G and A reach a standard agreement that each category of the categorical variables is a distinct binary feature to be managed inside the GAN framework. It is carried out with a multi-layer perceptron in the case of the generating function. The principal objective associated with this process will hence be to generate synthetic data that mimics the features of the original adult sample statistically.

In the meantime, the discriminator, which is also an MLP, plays an essential role in establishing the difference between the joint distribution of the generated observations and the actual differential private data. The use of Leaky ReLU activations[28] in all layers of the critic helps it classify genuine and synthetic data instances better, correcting the generator's output. Throughout Training, the discriminator measures what is called the Kullback–Leibler divergence, which is used to compare actual data distribution to the generated ones. This measure determines how close the synthetic data is to the origination data distribution. Through this, the generator aims to minimize this KL divergence and generate more accurate and valuable synthetic data outputs. The described approach implies the fine-tuning of the design to maintain the necessary amount of information extracted from the initial data while also providing high-quality synthesized data, thus meeting the objectives of this enhancing privacy project. The data employed is (i) the Census Income dataset from the UC machine learning repository with an attribute count of 15. The census income dataset is characterized by 48,843 tuples, which are described by fifteen characteristics. 60,561 are structurally used as training sets, with the rest of the records used as tests.

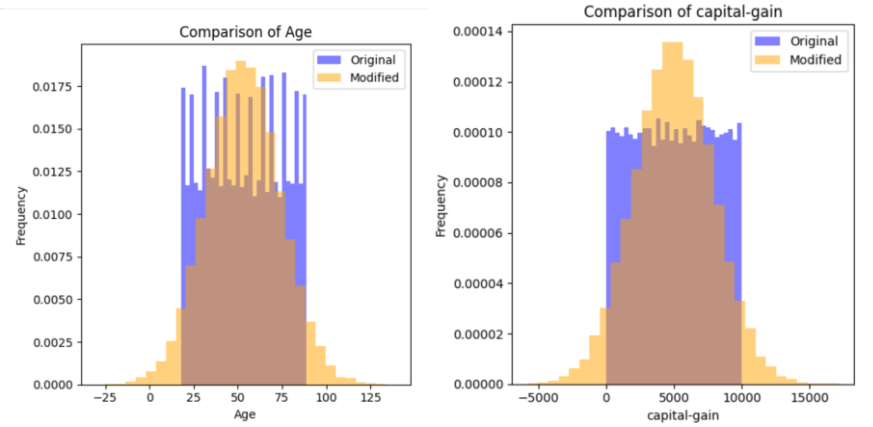
Simulation Setup: The experiments were performed using Google Colab with the intention of using Python 3 for all the scripts. 7, TensorFlow 2. 5, and Keras 2. 4. The adult dataset was preprocessed, where the numerical features were scaled, and the categorical features were encoded using one-hot encoding. SAs for patients were first processed by differential privacy methods to add noise to the quasi-identifiers (QI) before they were fed to the GAN model. For the generator and the discriminator network, a multi-layer perceptron fully connected network was used, and both of them comprise 3 dense layers, each with Leaky ReLU activation function for all layers except the last one, which has the sigmoid function of the discriminator network. Optimizations for both networks were done with the help of Adam's optimizer with a learning rate of 0. 001. Here, Training was done on 200 epochs with 128 batch sizes. Further, the generator was trained to minimize the "KL divergence" to the target distribution. In contrast, the discriminator was trained to maximize the probability that accurate data and created data have different distributions. Stabilization of the Training was done using gradient penalties. The evaluation that was carried out entailed the divergence between two distributions through the use of KL divergence and a general assessment of visual quality of synthetic data[30].

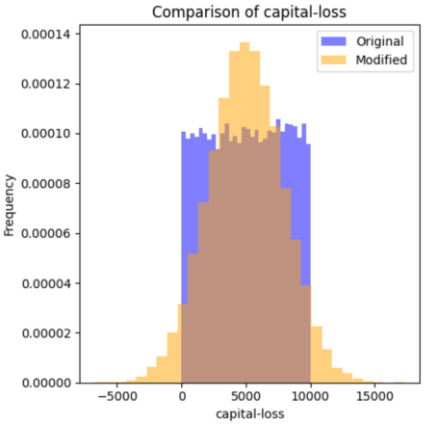
The implementation consisted of numerous steps: On the basis of the above criterion, the Adult Census Income dataset containing 15 variables was selected, and out of these variables, nine were categorical, and six were continuous. The processes of vertical

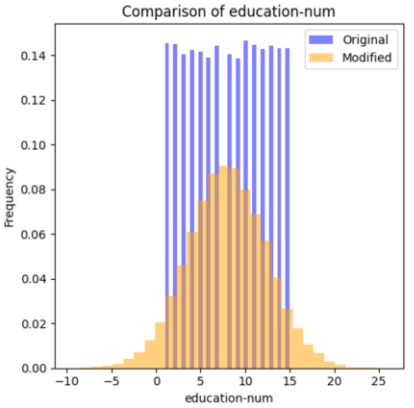
data fragmentation for QI and SA attributes were conducted with the help of algorithms. The following QI characteristics were subjected to the indicated μ-Gaussian DP: Nonstandard (ε, δ)-DP constraint was not applied, but the f-DP constraint, where f stands for the tightest trade-off function involved in the estimation. This framework, which is even stronger than (ε, δ)-DP, provides more privacy by allowing δ=0. Theoretical framework The presented work builds upon two general notions from the field of differential privacy: D 001 and the minimum ε for (ε, δ)-DP compliance. From the fragments of f-DP and (ε, δ)-DP, it can be shown that μ-GDP implies (ε, μ)-DP for any $\varepsilon \geq 0$, which includes a more accurate privacy bound than existing methods. , gauge the effect of privacy budget rates (ε, δ) on data correctness, Gaussian and Laplacian noise were added to the QI. (iii) The QI was contaminated with Gaussian and Laplacian noise in an attempt to determine the relationship between privacy budget rates (ε, δ) and data correctness.

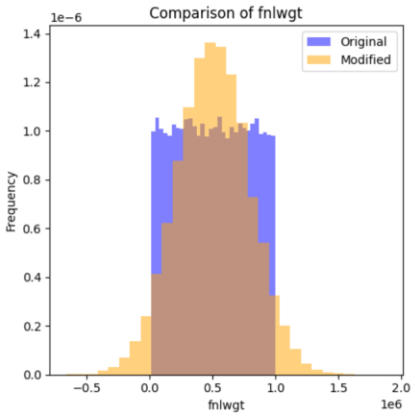
The changes that we make to the dataset can be observed to have led to shifts despite the fact that the distributions of the attributes are the same. In general, the frequency and the range of numerical properties may change but do not upset the general shape of the property. Whereas the frequencies of some specific categories have been changed, the distributions of categorical traits in general are analogous. These are the ones that display characteristics relating to different things. **Fig.6** provides a comprehensive comparison of all attributes.

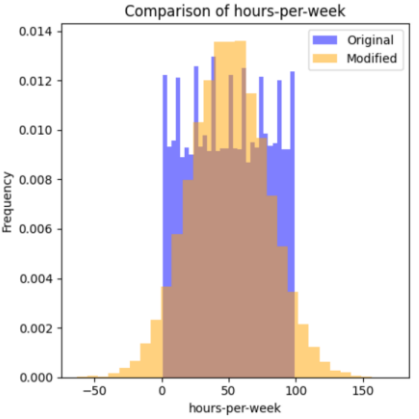
Fig.6: Comparison of all attributes in dataset to modified data

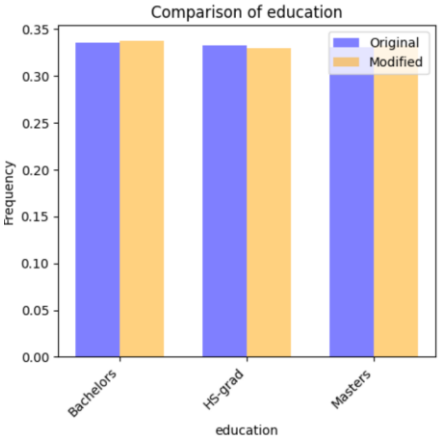


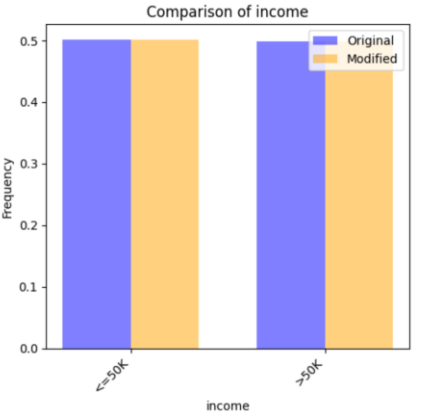

Comparison of capital-loss
Original
Modified
Frequency
capital-loss

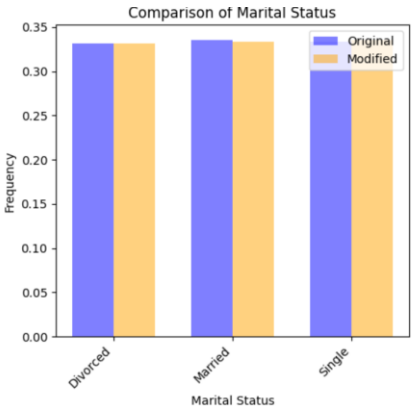

Comparison of education-num
Original
Modified
Frequency
education-num

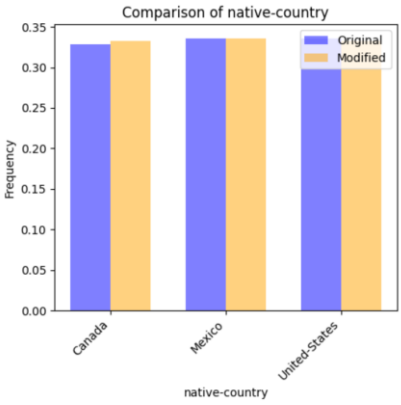

Comparison of fnlwgt
1e−6
Original
Modified
Frequency
fnlwgt
1e6

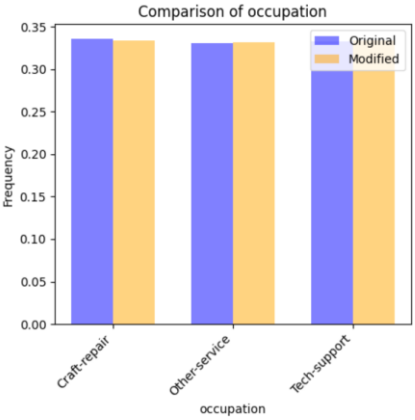

Comparison of hours-per-week
Original
Modified
Frequency
hours-per-week

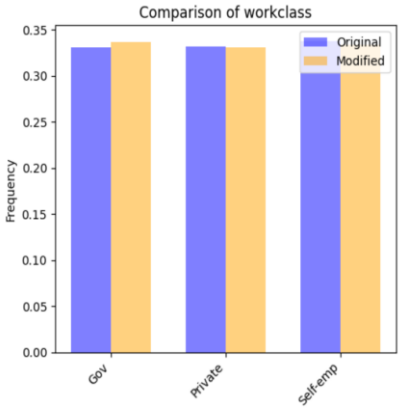

Comparison of education
Original
Modified
Frequency
Bachelors
HS-grad
Masters
education

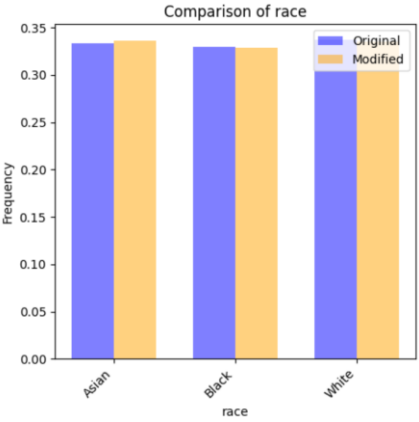

Comparison of income
Original
Modified
Frequency
<=50K
>50K
income

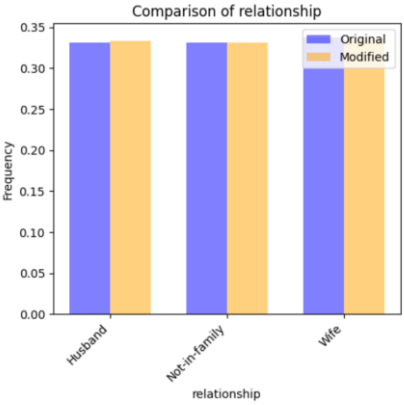

Comparison of Marital Status
Original
Modified
Frequency
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Divorced
Married
Single
Marital Status


Comparison of native-country
Original
Modified
Frequency
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Canada
Mexico
United-States
native-country


Comparison of occupation
Original
Modified
Frequency
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Craft-repair
Other-service
Tech-support
occupation


Comparison of workclass
Original
Modified
Frequency
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Gov
Private
Self-emp


Comparison of race
Original
Modified
Frequency
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Asian
Black
White
race


Comparison of relationship
Original
Modified
Frequency
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Husband
Not-in-family
Wife
relationship

5 Evaluation Metrics: Accuracy, Precision, Recall, and F1 Score

It is essential to evaluate multiple metrics that represent distinct facets of model performance in order to evaluate the efficacy of data privacy strategies. **Accuracy** measures the correctness of a model's predictions. It is defined as the ratio of correctly predicted instances to the total number of instances Accuarcy is as evaluated as in Eq.(11)

$$\text{Accuracy} = \frac{Number\ of\ correct\ Predictions}{Total\ Number\ of\ predictions} \tag{11}$$

Precision indicates the proportion of true positive predictions among all positive predictions made by the model. High precision ensures that when the model predicts a positive outcome, it is likely to be correct. Precision is as evaluated as in Eq.(12)

$$\text{Precision} = \frac{True\ Positives}{True\ Positives+False\ Positives} \tag{12}$$

Recall (or Sensitivity) measures the proportion of actual positives that were correctly identified by the model. High recall is crucial in applications where missing a positive instance could have serious consequences as evaluated as in Eq.(13)

$$\text{Recall} = \frac{True\ Positives}{True\ Positives+False\ Negetives} \tag{13}$$

F1 Score is the harmonic mean of precision and recall, providing a single metric that balances both aspects. It is especially useful when dealing with imbalanced datasets where neither precision nor recall alone is sufficient. F1 Score is as evaluated as in Eq.(14)

$$\text{F1 Score} = 2.\frac{Precision.Recall}{Precision+Recall} \tag{14}$$

We examined the differences in performance between Laplacian-modified data and Gaussian-perturbed data. All measures showed poorer results for the Gaussian-perturbed data, with averages for precision, recall, and F1 scores of about 50%. This implies that although Gaussian noise maintains certain aspects of the data, it is not as successful in preserving the data utility as Laplacian noise. Laplacian modification yielded far greater results, averaging 82% in F1, recall, and precision. This suggests that Laplacian noise increases overall data quality and utility in addition to improving the model's ability to correctly detect positive cases. The enhanced performance measurements demonstrate how well Laplacian noise balances data value and privacy. The modified dataset achieves an accuracy of 85.2%, precision of 85.75% , recall of 79.9% and F1 Score 82.8% reflecting a high similarity to the original data while incorporating privacy-preserving modifications. This accuracy indicates that the modifications balance privacy and data utility effectively, preserving the dataset's integrity. **Fig.7** compares the accuracy of data modified with Laplacian and Gaussian noise

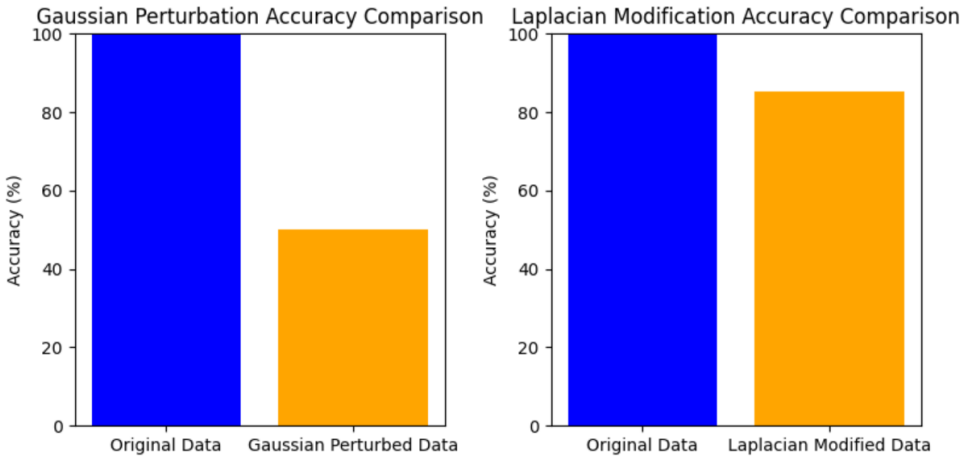


Fig.7: Accuracy of model based on Gaussian and Laplacian noise

It was established that its recognition capability was about 48 percent. The Gaussian data is evident to have a much lower accuracy than the presented original data as it appears in a totally different level of the bar chart, only 1%. Indeed, the Laplacian-modified data proposed to enhance privacy while compromising the usefulness responds with a much higher accuracy of 85.2%. Therefore, the Laplacian noise is more appropriate than the Gaussian perturbation for, say, the high precision and data usefulness required in certain instances. Fig.8 illustrates the overall accuracy of the modified data, summarizing the effectiveness of the Fortified-GAN approach.

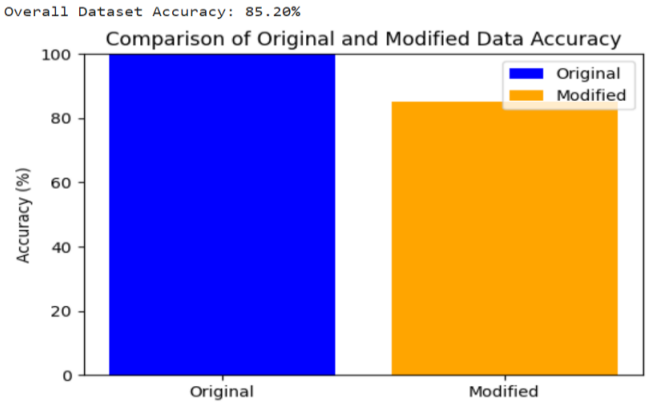


Fig.8: Accuracy of proposed model

6 Conclusion

The present paper suggests developing a methodology based on a variety of machine learning techniques for anonymization of the Adult Census Income dataset with equal concern for accuracy and privacy. As for the QI features, they were found and improved with the help of Gaussian and Laplacian noise, and GANs proved to be relatively efficient in creating synthetic data matching the primary dataset in terms of statistical characteristics. Through generator and discriminator networks, models could be trained, and synthetic data generated would provide high-quality results while protecting users' privacy.

Future work will focus on integrating feature engineering methods with differential privacy (DP) to improve the privacy-utility ratio, specifically deploying with epsilon-differential privacy (ε-DP) methodologies. Developing new techniques for handling categorical data in GANs, addressing the current limitations. Additionally, the impact of different privacy budgets on model performance will be explored, optimizing the balance between data privacy and utility. Expanding this approach to other datasets will also be pursued, further validating the robustness and adaptability of the proposed methods.

References

1.Elliot, M.J., Manning, A., Mayes, K., Gurd, J., Bane, M.: SUDA: A Program for Detecting Special Unique. In: Proceedings of UNECE Work Session on Statistical Data Confidentiality (2005).

2.Lodha, S., Thomas, D.: Probabilistic anonymity https://doi.org/10.1007/978-3-540-78478-4_4 (2008).

3.Adult Census Income Dataset from UCI Machine learning repository https://archive.ics.uci.edu/dataset/2/adult

4.Lichman, M.: UCI Machine Learning Repository. http://archive.uci.edu/ml (2013).

5.Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. In: Proceedings of the Advances in Neural Information Processing Systems 2014. https://papers.nips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

6. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings (2014). https://arxiv.org/abs/1312.6114

7. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput.Sci. (2013). https://doi.org/10.1561/0400000042

8.Dwork, C., Rothblum, G.N., Vadhan, S.: Boosting and differential privacy. In: Proceedings - Annual IEEE Symposium on Foundations of Computer Science, FOCS

(2010).

9.Dwork, C.: Differential Privacy: A Survey of Results. Theory and Applications of Models of Computation. Springer, Berlin (2008).

10.Dwork, C., Kenthapadi, K., McSherry, F., et al.: Our data, ourselves: Privacy via distributed noise generation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (2006).

11. Loukides, G., Liagouris, J., Gkoulalas-Divanis, A., Terrovitis, M.: Disassociation for electronic health record privacy. J. Biomed. Inf. (2014). https://doi.org/10.1016/j.jbi.2014.05.009

12. Acs, G., Melis, L., Castelluccia, C., De Cristofaro, E.: Differentially private mixture of generative neural networks. IEEE Trans Knowl Data Eng (2019). https://doi.org/10.1109/TKDE.2018.2855136

13. Lee, H., Kim, S., Kim, J.W., Chung, Y.D.: Utility-preserving anonymization for health data publishing. BMC Med. Inf. Decis. Mak. (2017). https://doi.org/10.1186/s12911-017-0499-0

14. Hitaj, B., Ateniese, G., Perez-Cruz, F.: Deep Models under the GAN: Information leakage from collaborative deep learning. In: Proceedings of the ACM Conference on Computer and Communications Security (2017). https://dl.acm.org/doi/10.1145/3133956.3134012

15. Fredrikson, M., Lantz, E., Jha, S., et al.: Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: Proceedings of the 23rd USENIX Security Symposium (2014).

16.Manning, A.M., Haglin, D.J.: A new algorithm for finding minimal sample uniques for use in statistical disclosure assessment. In: Proceedings of the IEEE International Conference on Data Mining, ICDM (2005).

17.Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. In: Proceedings of the 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings (2017).

18. Zhang, J., Cormode, G., Procopiuc, C.M., et al.: PrivBayes: Private data release via Bayesian networks. In: Proceedings of the ACM SIGMOD International Conference on Management of Data (2014). https://dl.acm.org/doi/10.1145/2588555.2593621

19.Motwani, R., Xu, Y.: Efficient Algorithms for Masking and Finding Quasi-Identifiers. VLDB ’07 (2007).

20.Dong, J., Roth, A., Su, W.J.: Gaussian differential privacy(2019).

21. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief

nets. Neural Comput. (2006). https://doi.org/10.1162/neco.2006.18.7.1527

22. Lu, Y., Sinnott, R.O., Verspoor, K.: A semantic-based k-anonymity scheme for health record linkage. In: Studies in Health Technology and Informatics (2017).

23. Sweeney, L.: k-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. (2002). https://doi.org/10.1142/S0218488502001648

24. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: ℓ-Diversity: privacy beyond k-anonymity. In: Proceedings of the International Conference on Data Engineering(2006)
https://www.cis.upenn.edu/~mkearns/teaching/cgtoc/machanavajjhala06ldiversity.pdf

25. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership Inference Attacks Against Machine Learning Models. In: Proceedings of the IEEE Symposium on Security and Privacy (2017). https://ieeexplore.ieee.org/document/7958568

26.M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. on Computer and Communications Security (CCS)*, 2015, pp. 1322-1333. doi: 10.1145/2810103.2813677.

27. Beaulieu-Jones, B.K., Wu, Z.S., Williams, C., Lee, R., Bhavnani, S.P., Byrd, J.B., Greene, C.S.: Privacy-preserving generative deep neural networks support clinical data sharing. Circulation 12(7), e005122 (2019). https://www.ahajournals.org/doi/full/10.1161/CIRCOUTCOMES.119.00512

28.J. Shlens, "A tutorial on the Kullback-Leibler divergence," arXiv preprint arXiv:1404.2000, 2014. [Online]. Available: https://arxiv.org/abs/1404.2000.

29.A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, 2013.

30. Abay, N.C., Zhou, Y., Kantarcioglu, M., et al.: Privacy preserving synthetic data release using deep learning. In: Lecture Notes in Computer Science (2019). https://link.springer.com/chapter/10.1007/978-3-030-16145-3_18