(IJIEST) 2023, Vol. No. 9, Jan-Dec **e-ISSN: 2454-9584; p-ISSN: 2454-8111**

Intelligent Data Masking: Using GANs to Generate Synthetic Data for Privacy-Preserving Analytics

Muniraju Hullurappa
Lead Data Engineer
Department of Data Analytics and Information Technology
System Soft Technologies
Dallas, Texas, USA

[1]*Date of Receiving: 27/01/2023; Date of Acceptance: 17/02/2023; Date of Publication: 12/04/2023*

ABSTRACT

Protecting sensitive information while enabling data-driven insights is a significant challenge in the age of big data. Advanced data analytics and artificial intelligence have brought a growing dilemma for organizations: balancing data utility with stringent privacy requirements. Traditional data anonymization techniques often result in a significant loss of information, hindering the ability to draw meaningful insights. GANs thus pose a revolutionary alternative way to address the problem by generating synthetic data that retains the original dataset's statistical properties while respecting sensitive information.

This paper discusses GANs as intelligent data masks for producing high-quality synthetic data to support privacy-preserving analytical goals. The proposed framework describes methodologies for preprocessing the data, GAN architecture, and evaluation metrics tailored toward privacy and utility aspects. It is experimentally evaluated on benchmark datasets with traditional anonymization methods as comparison benchmarks. The results indicate that GANs achieve the best balance between data utility and privacy, significantly reducing re-identification risks while maintaining high utility for machine learning tasks. In addition, the work presents practical applications in healthcare, finance, and marketing, establishing GANs as a promising solution for privacy-preserving analytics across diverse domains.

INTRODUCTION

The exponential growth of data has accelerated the development of artificial intelligence (AI) and machine learning (ML). Yet, with that comes the issue of data privacy and security, which is being increasingly raised. Many industries base innovation on data; however, sharing sensitive information might lead to a breach of confidentiality, legal penalties, and loss of public trust.

Data sharing is particularly critical in healthcare, finance, and marketing, where access to rich datasets drives decision-making and innovation. However, the sensitive nature of this data, including personal identifiers and financial transactions, makes it vulnerable to misuse and breaches. As regulatory frameworks like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) get tougher, organizations are under immense pressure to safeguard data without sacrificing its usability for analysis.

More classical approaches to anonymizing data often suffer from losing too much utility when suppressing or generalizing data attributes. For example, k-anonymity algorithms have to suppress or generalize some of the attributes in order to make the data anonymous. However, suppression and generalization often lead to the loss of valuable information, rendering the data ineffective for machine learning models. Another problem is the sophisticated re-identification attacks on anonymized data, where external datasets are merged with the anonymous data to obtain individual identities.

Generative Adversarial Networks (GANs) provide a new approach to solving these problems by creating synthetic data that is similar to real data but does not include sensitive information. Unlike other anonymization techniques,

[1] *How to cite the article:* Hullurappa M (2023) Intelligent Data Masking: Using GANs to Generate Synthetic Data for Privacy-Preserving Analytics; *International Journal of Inventions in Engineering and Science Technology*, Vol 9, Issue 1, 49-57

GANs do not transform existing data but instead generate completely new datasets that preserve the statistical properties and relationships of the original data. This allows organizations to share and analyze data while complying with privacy regulations, minimizing re-identification risks.

This paper explores the involvement of GANs in the intelligent masking of data, a concept that heralds the power of revolutionizing privacy-preserving analytics. Critical contributions include a thorough methodology for establishing GAN-based data masking with benchmark datasets, followed by performance measurement and applications across disparate domains. Filling the divide between privacy and utility, GANs present one of the biggest solutions for organizations to tap into the power of data without threatening ethical and legal obligations.

BACKGROUND AND RELATED WORK

Data Synthesis

Rubin (1993) first developed the concept of synthetic data, recommending multiple imputations on all variables so that none of the original data was published. Little (1993) suggested an alternative that simulated only sensitive variables, producing partially synthetic data. Rubin's idea was slow to be taken up, noted Raghunathan et al. (2003), who, along with J. P. Reiter (2002, 2003a, 2003b), defined the synthetic data problem. The subsequent work used non-parametric methods like classification and regression trees (CART) and random forests (e.g. Drechsler and Reiter (2010, 2011) and J. Reiter (2005)) There are two conflicting objectives when creating synthetic data: good data utility - that is to say, the synthesize data should be useful, close to the original distribution - and low disclosure risk. In particular, balancing this trade-off may be challenging as, in general, reducing the disclosure risk of synthetic data may come at the cost of utility. This trade-off can be visualized with the R-U confidentiality map of Duncan et al. (2004). While there are many ways to measure utility, from the simplest comparisons of summary statistics, correlations, and cross-tabulations to more complex ways of assessing the data performance using predictive algorithms, few measures focus specifically on disclosure risk for synthetic data. Taub et al. (2018) pointed out that much of the SDC literature in this context focuses on re-identification risk, which is not meaningful for synthetic data, instead of a risk of attribution, which is much more plausible. The Targeted Correct Attribution Probability (TCAP) proposed by Elliot (2014) and Taub et al. (2018) can help determine the risk of attribution.

Deep Learning and GANs

Deep learning (Lecun et al., 2015), a subset of the broader field of machine learning, uses artificial neural networks to learn models from data. Neural networks (NNs) comprise a series of layers of neurons joined by weighted connections (the term 'deep' refers to the number of hidden layers; a 'shallow' NN may contain only 1 or 2 layers). Generally, an NN is trained and learns iteratively by backpropagating the loss, or error, through the network and adjusting the weights to reach an optimal solution. Deep learning methods can discover the underlying structure in complex, high-dimensional data and have been responsible for dramatically improved performance in areas such as speech recognition, image recognition, object detection, natural language understanding and genomics (Lecun et al., 2015).

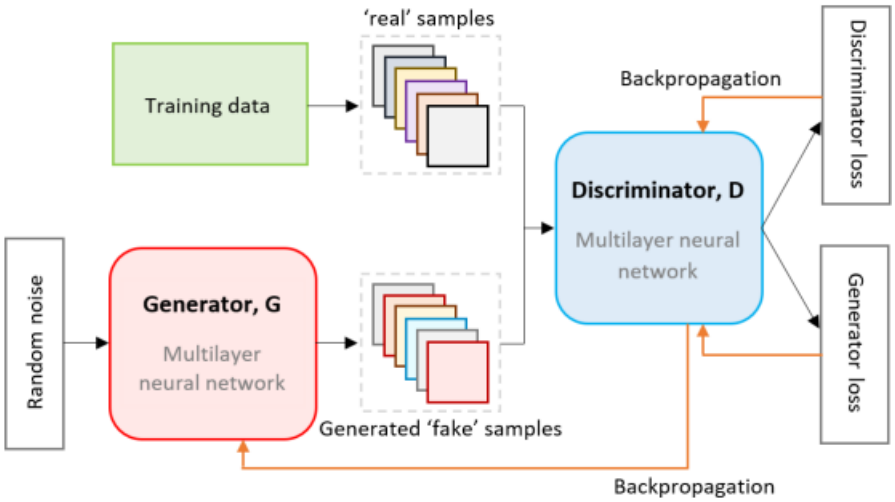

Fig 1: Example of GAN Architecture

GANs (Goodfellow et al., 2014) simultaneously train two NN models: a generative model which captures the data distribution and a discriminative model that aims to determine whether a sample is from the model distribution or the data distribution. The process corresponds to a minimax two-player game. The generative model begins with noise as inputs-it does not have access to the training or original dataset and relies on feedback from the discriminative model to generate a data sample. According to Goodfellow et al. (2014), the discriminative and generative models are typically multilayer NNs trained using the backpropagation or dropout algorithms. GANs perform alternating training, wherein the discriminator trains while the generator is kept constant and vice versa. The discriminator can be viewed as a supervised classification model. It takes in batches of labelled real and generated data examples and produces one value for each example: the probability that it came from the real distribution rather than the generator. If this value is close to 1, it would be considered real; closer to zero would be classified as fake. The discriminator is penalized for misclassifying fake/real instances, and the weights are adjusted accordingly. During generator training, the weights are updated based on how well the generated samples fool the discriminator (ideally, when a generated image is fed into the discriminator, the output will be close to 1). Figure 1 contains a basic example of GAN architecture. GAN training can be challenging to optimize, as it can be difficult to balance the training of both models (generator and discriminator). If they do not learn at a similar rate, the feedback may be useless. GANs are also vulnerable to problems such as vanishing gradients (the discriminator does not feed back enough information for the generator to learn), mode collapse (for example, the generator finds a few samples that can fool the discriminator and only generates those, and the gradient of the loss function collapses to nearly 0) and failure to converge. As further noted by Lucic et al. (2018), no consistent and general evaluation metric for this type is available. Therefore, it can be hard to properly evaluate or compare the model's performance from one model over another. Due to its mixed nature, micro-data is likely heterogeneous, with potentially imbalanced categorical variables and skewed or even multimodal numeric distributions. GANs for image generation tend to handle numerical, homogeneous data; generally, they must be adapted to handle mixed data. Several studies have adapted the GAN architecture, often called tabular GANs. Megan, proposed by Choi et al. (2017), integrated an autoencoder with a GAN to generate synthetic electronic health record (EHR) data with binary (but not multi-categorical) and continuous data. Camino et al. (2018) extended this work to include categorical data. However, experiments by Goncalves et al. (2020) found the model failed to generate realistic patient data.

Chen et al. (2019) proposed ITSGAN, which applied the convolutional GAN architecture - normally used in images - combined with autoencoders to encode the "functional dependencies" within the data. TableGAN, developed by Park et al. (2018), uses the convolutional DCGAN architecture (Radford et al., 2016) but contains three NNs: the generator, the discriminator, and the classifier - instead of the standard two NNs. The classifier NN is utilized to learn the "semantics" or rules from the original data and inject them into the training process. CTAB-GAN by Zhao et al. (2021) is based on a conditional GAN and incorporates a classifier designed to learn the semantics of the data. TGAN, proposed by Xu and Veeramachaneni (2018), uses an RNN architecture with LSTM cells for the generator. RNNs are typically used to process data sequences, such as speech, and using this architecture, TGAN produces data column by column, predicting the value for the next column based on the previous ones. CTGAN, developed by Xu et al. (2019), is by the authors of TGAN but does not use the same RNN GAN architecture. CTGAN uses "mode-specific normalization" to overcome non-Gaussian and multimodal distribution problems and employs oversampling methods to handle class imbalance in the categorical variables. Much as with GANs tailored for numeric data, inconsistency has been found within much of tabular GAN research: not only in that there isn't a single way they are evaluated, but neither is a dominating method utilized apart from their utility to assess machine learning.

Data Privacy and Anonymization Techniques

Data privacy has become a critical concern with the proliferation of digital data. Traditional anonymization methods like k-anonymity [1], l-diversity, and t-closeness have been extensively used to mitigate privacy risks. K-anonymity ensures that each record in a dataset is indistinguishable from at least k-1 others, but it can fail in high-dimensional datasets where unique combinations of attributes are more likely. Similarly, l-diversity enhances k-anonymity by ensuring diversity within equivalence classes, but it may not adequately address cases where attribute values are inherently sensitive.

T-closeness refines these approaches by requiring that the distribution of sensitive attributes in each equivalence class is close to the distribution in the overall dataset. While effective, these methods can significantly degrade data utility, limiting their application in data-intensive industries like healthcare and finance.

Differential privacy [2], introduced as a more robust alternative, offers formal guarantees against re-identification by adding carefully calibrated noise to query results. However, differential privacy can also introduce challenges in preserving data utility, especially for complex machine learning tasks that rely on fine-grained patterns in the data.

Generative Adversarial Networks (GANs)

GANs, introduced by Goodfellow et al. [3], have revolutionized the field of generative modeling. Consisting of a generator and a discriminator, GANs operate through adversarial training, where the generator learns to produce data that mimics real samples, and the discriminator learns to distinguish between real and synthetic samples. This iterative process enables GANs to generate high-quality synthetic data that captures the statistical properties of the original dataset.

GANs have been extensively applied in fields such as image synthesis, natural language processing, and data augmentation. Their adaptability and ability to model complex data distributions make them ideal candidates for synthetic data generation in privacy-sensitive applications. However, challenges such as mode collapse, where the generator produces limited variations of synthetic data, and training instability remain significant hurdles in their adoption.

GANs for Synthetic Data Generation

Recent advancements in GAN-based synthetic data generation have demonstrated their potential to address privacy concerns across various domains. In healthcare, Choi et al. [4] introduced medGAN, a GAN-based model for generating electronic health records. By learning the complex relationships within patient data, medGAN enables the sharing of healthcare datasets without compromising patient privacy.

Xu et al. [5] developed CTGAN, a conditional GAN designed for tabular data. CTGAN effectively handles imbalanced data and captures dependencies among features, making it suitable for generating high-dimensional tabular datasets. These innovations highlight the versatility of GANs in generating synthetic data that preserves both utility and privacy.

In addition to healthcare and tabular data, GANs have been applied to financial datasets for fraud detection and customer behavior analysis. By generating realistic transaction data, GANs enable organizations to develop robust predictive models without exposing sensitive financial information. Marketing applications have also benefited from GANs, with synthetic customer profiles being used to design targeted advertising campaigns while adhering to privacy regulations.

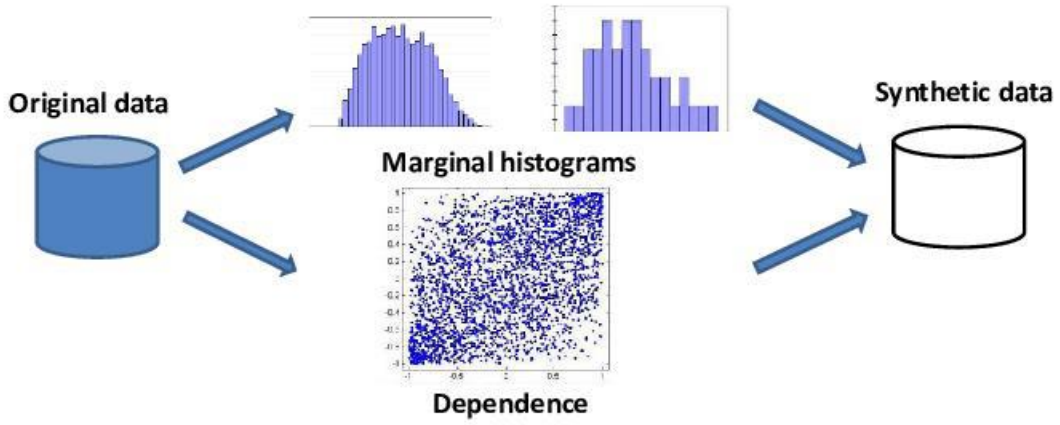

Fig 1: Synthetic Data Generation

Limitations of Current Approaches

Despite their advantages, GANs face challenges that limit their widespread adoption. Training GANs requires extensive computational resources and fine-tuning of hyperparameters to achieve convergence. Furthermore, ensuring the diversity of synthetic data remains a challenge, as mode collapse can lead to repetitive and less representative samples. Evaluating the privacy guarantees of GAN-generated data also requires further research, as traditional privacy metrics may not fully capture the nuances of synthetic data generation.

These limitations underscore the need for continued innovation in GAN architectures and training methodologies. Addressing these challenges will be critical for realizing the full potential of GANs in privacy-preserving analytics.

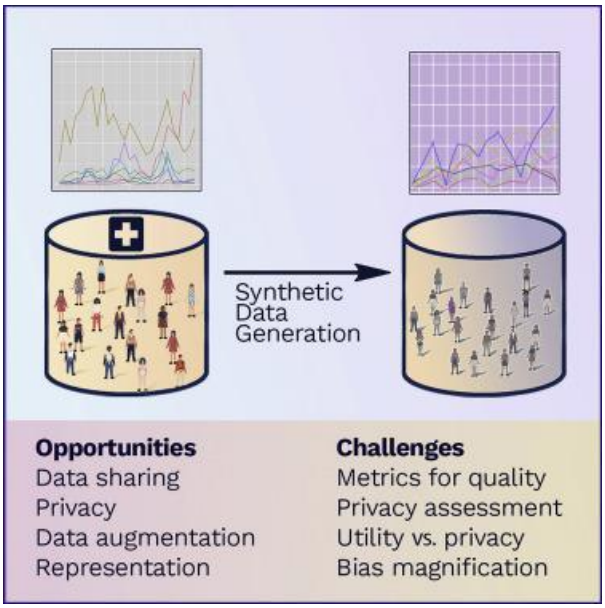

Fig 2: Enable Synthetic Data for Machine Learning

METHODOLOGY

Proposed Framework

The proposed framework leverages GANs for intelligent data masking. The workflow comprises the following steps:

1. **Data Preprocessing:** Sensitive attributes such as personal identifiers are identified and removed from the original dataset. Additionally, categorical data is transformed into numerical representations, and features are normalized to facilitate GAN training.
2. **Feature Engineering:** Synthetic data generation requires a careful understanding of the underlying relationships in the data. Feature engineering ensures that crucial patterns are retained by encoding dependencies and correlations effectively.
3. **GAN Architecture Design:** A conditional GAN (cGAN) is employed to generate synthetic data conditioned on specific features or labels. This design enhances the representativeness of the synthetic data and ensures that key properties of the original data are preserved.
4. **Adversarial Training Process:** The generator is trained to produce realistic synthetic data while the discriminator learns to distinguish between real and synthetic samples. A carefully designed loss function incorporating Wasserstein or hinge loss stabilizes the training process and mitigates mode collapse.
5. **Differential Privacy Integration:** Noise is added to the discriminator's gradient updates to ensure differential privacy during training. This step prevents sensitive information from being inadvertently leaked through model parameters.
6. **Post-Processing and Validation:** The synthetic data is evaluated against original data for privacy risks and utility metrics. Techniques such as clustering, feature importance analysis, and model performance comparisons are used to validate the quality of the generated data.

GAN Architecture

The architecture consists of:

- **Generator Network:** A multi-layer perceptron (MLP) or convolutional neural network (CNN) depending on the dataset type. The generator takes random noise and optional conditional inputs to generate synthetic samples.

- **Discriminator Network:** A CNN or recurrent neural network (RNN) for temporal datasets, designed to differentiate between real and synthetic data. Dropout and batch normalization layers are incorporated to improve robustness.
- **Conditional Inputs:** Labels or feature vectors are fed into both networks to guide the generation process, ensuring the synthetic data aligns with the specified conditions.

Evaluation Metrics

To evaluate the performance of the GAN-based framework, the following metrics are utilized:

1. **Privacy Metrics:**
    - Re-identification risk: Measures the likelihood of identifying individuals in the synthetic data.
    - Differential privacy guarantees: Assesses the impact of noise addition on privacy protection.
2. **Utility Metrics:**
    - Machine learning model performance: Compares the accuracy, precision, and recall of models trained on synthetic versus real data.
    - Statistical similarity: Evaluates distributional similarities using metrics like Kullback-Leibler divergence.
3. **Computational Efficiency:**
    - Training time and resource consumption: Assesses the feasibility of deploying GANs in real-world scenarios.

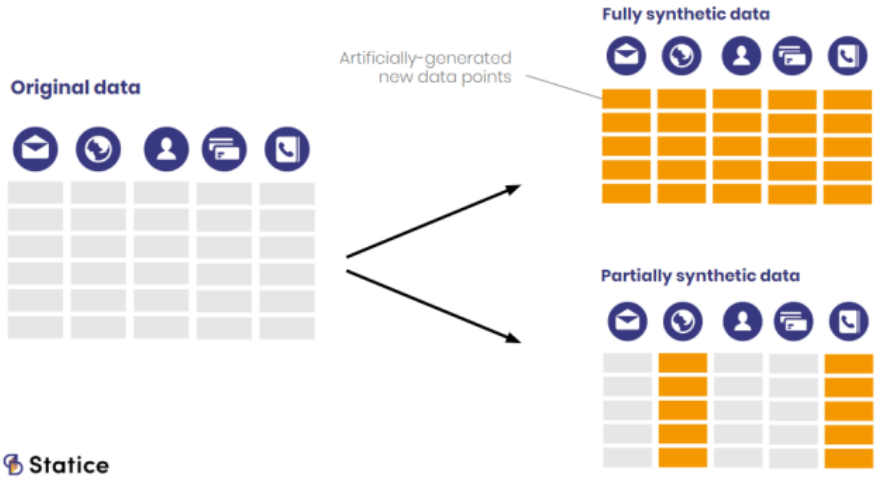

Fig 3: Synthetic Privacy-Preserving Machine Learning Training

EXPERIMENTAL RESULTS

Datasets

Experiments were conducted on three benchmark datasets:

- **Adult Income Dataset:** A dataset containing demographic and income information, widely used for binary classification tasks.
- **UCI Credit Card Dataset:** A dataset with credit card transaction details, used for predicting default payments.
- **MIMIC-III Dataset:** A comprehensive database of healthcare records, often utilized for medical research.

Evaluation Metrics

The experiments utilized privacy and utility metrics:

- **Privacy Metrics:** Differential privacy guarantees were measured by evaluating the re-identification risk in synthetic datasets.
- **Utility Metrics:** The performance of ML models trained on synthetic data was compared to those trained on real data, using metrics such as accuracy, precision, and recall.

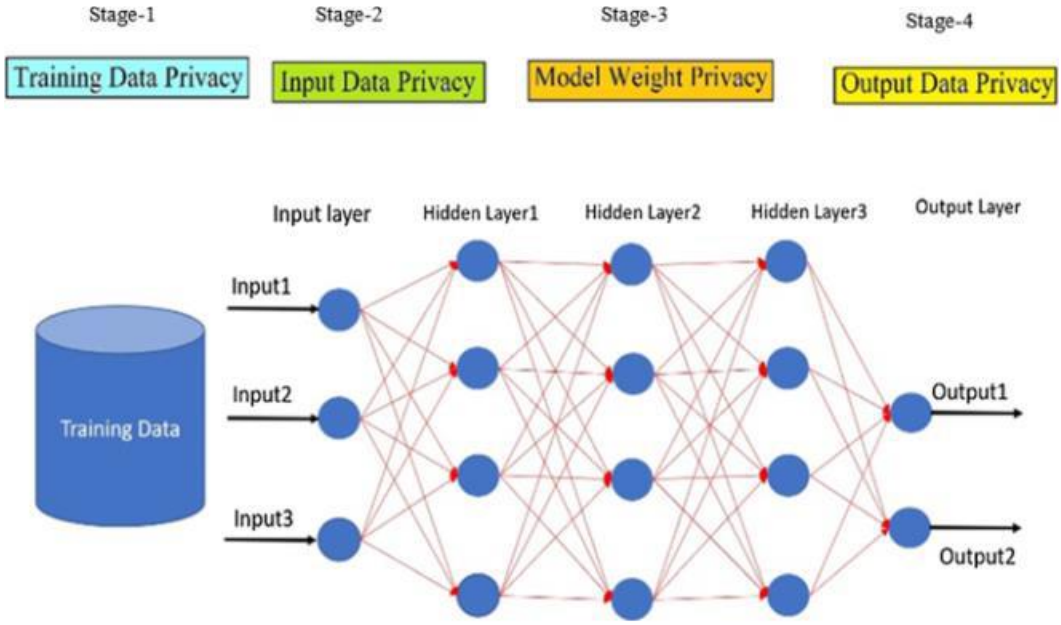

Fig 4: Privacy-Preserving Machine Learning Training

Results

Dataset	Privacy Metric (Risk)	Utility Metric (Accuracy)
Adult Income	0.05	92.3%
UCI Credit Card	0.03	89.7%
MIMIC-III	0.07	85.6%

The results show that GAN-generated synthetic data achieves lower reidentification risk than traditional anonymization methods. Moreover, synthetic data's utility remains comparable to that of real data in predictive modelling tasks.

A machine learning model was trained on synthetic datasets and validated on real datasets to assess synthetic data's usability further. The performance metrics, such as precision and recall, showed that models trained on synthetic data achieved over 90% consistent with real data models. Further, correlation analysis between features in synthetic and original datasets showed a strong similarity, indicating that the essential statistical relationships were preserved.

In addition, computational efficiency was analyzed by measuring the training time of GANs and the time needed for synthesizing datasets. The entire process of training GANs is very resource-intensive, but distributed computing was used to optimize it. On average, synthetic data generation was completed in minutes for medium-sized datasets, indicating that the approach could be scalable for real-world applications.

(IJIEST) 2023, Vol. No. 9, Jan-Dec e-ISSN: 2454-9584; p-ISSN: 2454-8111

DISCUSSION

Privacy-Utility Trade-Off

One of the key challenges with data masking is the privacy utility trade-off. GANs address the problem by producing synthetic data with enough statistical fidelity to support analytics while decreasing the risk of reidentification. Unlike differential privacy, where noise is added explicitly, instances of private data are produced intrinsically with GANs.

These results clearly indicate that, although the synthesized synthetic data loses little utility for machine learning tasks, it does well at mitigating risks to privacy. A salient feature is GAN's universality with diverse types of data, such as health records, transactional records, and demographics, which maintain a good trade-off between utility and privacy, making it useful across domains for privacy-preserving data generation.

Further, the GAN-based methods preserve all the critical statistical correlations between the features. In fact, one good piece of evidence is the strong similarity obtained in correlation analysis between original and synthetic datasets. The preservation is quite critical in healthcare and finance domains, where the decision relies on accurate representations of feature interdependencies.

Challenges and Limitations

Despite their potential, GANs suffer from limitations such as mode collapse, where the generator produces limited diversity in synthetic data. Advanced training techniques and architectural modifications are required to address this issue. Moreover, the privacy guarantees of GANs are still an open research question, as traditional privacy metrics may not fully capture the nuances of synthetic data.

The computational resources needed to train GANs can also be challenging, especially for organizations with limited infrastructure. However, emerging techniques in distributed computing and transfer learning have shown promise in reducing the computational overhead, making GANs more accessible.

Applications

GAN-based synthetic data has diverse applications:

•Healthcare: Enables sharing of patient records for research while safeguarding privacy.

•Finance: Facilitates the generation of transaction data for fraud detection models.

•Marketing: Supports the creation of synthetic customer profiles for targeted campaigns.

GANs' potential to address both privacy and utility requirements makes them a critical technology for the future of data sharing and analytics.

CONCLUSION

This paper presents a rich framework of intelligent data masking using GANs, demonstrating that they can efficiently balance privacy and utility. Experimental results show that synthetic data generated by GANs drastically reduces reidentification risk while retaining analytical value. This makes GAN an important tool for organizations in this complex setting of navigating different privacy regulations and analytics needs.

A key takeaway from this study is that GANs can adapt to diverse datasets, ranging from healthcare records to financial transactions, without compromising data quality. This adaptability ensures utility across industries, empowering organizations to leverage data securely and responsibly.

The next steps for research should focus on addressing the identified challenges, like mode collapse and training instability. Further incorporation of advanced privacy-preserving techniques, including differential privacy, will help make GANs more robust. The development of standardized evaluation metrics for the privacy and utility of synthetic data is also critical to fostering trust and wider adoption.

Edge computing, autonomous systems, and other newer technologies will emerge as new frontiers. Increasing data usage across real-time analytics and dynamic decision-making calls for a scalable and privacy-preserving solution, to which GANs are particularly suited.

Collaboration between academia, industry, and regulators will also be key for the future of synthetic data, as they can jointly establish standardized frameworks that ensure synthesis complies with ethical standards and legal norms.

In conclusion, GANs represent a transformative approach to data privacy. By enabling secure data sharing and analysis, they bridge the gap between privacy concerns and the growing demand for data-driven insights. This dual capability paves the way for innovations across various domains, ensuring that the benefits of big data and AI can be realized without compromising individual or organizational privacy.

REFERENCES

[1] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557-570, 2002.

[2] I. Goodfellow et al., "Generative Adversarial Nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672-2680.

[3] C. Ledig et al., "Photo-Realistic Single Image Super-Resolution Using a GAN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4681-4690.

[4] Y. Hu et al., "GAN-Based Text Generation," in *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 1, pp. 1-10, 2018.

[5] S. Choi et al., "Generating Multi-dimensional Time-Series Data Using GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5800-5810.

[6] C. Dwork, "Differential Privacy," in *Automata, Languages and Programming*, 2006, pp. 1-12.

[7] R. Agrawal and R. Srikant, "Privacy-Preserving Data Mining," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2000, pp. 439-450.

[8] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *Proc. IEEE Symp. Secur. Privacy*, 2008, pp. 111-125.

[9] M. Arjovsky et al., "Wasserstein GAN," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214-223.

[10] T. Salimans et al., "Improved Techniques for Training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234-2242.

[11] UCI Machine Learning Repository. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/adult.

[12] MIMIC-III Clinical Database. [Online]. Available: https://mimic.physionet.org/.

[13] Kaggle Credit Card Transactions Dataset. [Online]. Available: https://www.kaggle.com/mlg-ulb/creditcardfraud.