INTERNATIONAL MULTIDISCIPLINARY BOOK SERIES
https://ibseries.com/
Security and Privacy in AI Systems, Edited By: Dr. Sunita Chaudhary, Dr. Joydeb Patra

IMBS, Vol. 03, No. 1, (2025)
Open Access. Online & Print
ISSN:

# CHAPTER 4
# SECURITY IMPLICATIONS OF SYNTHETIC DATA GENERATION: MEMBERSHIP INFERENCE AND MODEL LEAKAGE RISKS

Roshmi Paul
BALLB Programme, Brainware University

**Abstract**
Synthetic data generation has emerged as a widely adopted privacy-preserving technique in machine learning, particularly in sensitive domains such as healthcare, finance, and social sciences. By replacing real datasets with artificially generated samples, organizations aim to mitigate direct privacy risks. However, recent advances reveal that synthetic data is not inherently secure. This study critically examines the security implications of synthetic data generation, focusing on membership inference attacks (MIAs) and model leakage risks. Membership inference allows adversaries to determine whether specific records were part of the training dataset, thereby compromising privacy. Additionally, generative models such as GANs and VAEs may inadvertently memorize and reproduce sensitive patterns, leading to data leakage. Through a synthesis of recent empirical studies, this paper analyzes attack mechanisms, evaluates vulnerabilities across different synthetic data paradigms, and discusses mitigation strategies including differential privacy and regularization. The findings highlight that while fully synthetic data offers improved protection, partially synthetic and overfitted models remain highly vulnerable. The study concludes by proposing a risk-aware framework for balancing privacy and utility in synthetic data deployment.

**Keywords**
Synthetic Data, Membership Inference Attack, Model Leakage, Privacy Risk, Generative Models, Differential Privacy, Data Security, GANs, VAE, Machine Learning Security

## 1. Introduction

The increasing reliance on data-driven technologies has intensified concerns regarding data privacy and security. Synthetic data generation has been proposed as a solution to enable data sharing without exposing sensitive information. Unlike anonymization techniques, synthetic data attempts to replicate statistical properties without preserving direct identifiers.

However, emerging research demonstrates that synthetic data may still leak sensitive information due to model memorization and overfitting. Generative models can encode latent representations of real training samples, making them vulnerable to adversarial attacks.

One of the most critical threats is the membership inference attack (MIA), where an attacker determines whether a specific individual's data was used during model training. Studies show that such attacks can be highly effective, particularly when models overfit or when synthetic data closely resembles real samples (PMC).



ISBN (978-81-685212-9-2)

## 2. Background

### 2.1 Synthetic Data Generation

Synthetic data generation refers to the process of creating artificial datasets that replicate the statistical properties, patterns, and relationships of real-world data without directly exposing sensitive information. It has gained significant importance in domains where privacy, confidentiality, and regulatory compliance are critical, such as healthcare, finance, and social sciences.

**Definition and Concept**

Synthetic data is generated using computational models that learn the underlying distribution of a real dataset and then produce new data points drawn from that learned distribution. Unlike anonymization, which modifies existing data, synthetic data is newly created, aiming to preserve:

- Statistical similarity
- Structural relationships
- Feature correlations

At the same time, it attempts to minimize direct linkage to real individuals.

**Types of Synthetic Data**

Synthetic data can be broadly categorized into three types:

| Type | Description | Privacy Level | Utility |
|---|---|---|---|
| Fully Synthetic | Entire dataset is artificially generated | High | Moderate |
| Partially Synthetic | Some attributes replaced with synthetic values | Low–Medium | High |
| Hybrid Synthetic | Combination of real and synthetic records | Medium | High |

- Fully synthetic data provides stronger privacy but may lose fine-grained accuracy.
- Partially synthetic data retains higher utility but is more vulnerable to leakage and inference attacks.

**Techniques for Synthetic Data Generation**

**1. Generative Adversarial Networks (GANs)**

- GANs consist of two neural networks:
- Generator: Creates synthetic data
- Discriminator: Evaluates authenticity

They compete in a minimax game, improving realism over time. GANs are widely used for image, tabular, and medical data synthesis, but are prone to overfitting and memorization, increasing leakage risks.

**2. Variational Autoencoders (VAEs)**

VAEs encode input data into a latent space and then reconstruct it.

**Advantages:**

- Stable training
- Better generalization

**Limitation:**

May produce less sharp or realistic outputs

VAEs reduce memorization compared to GANs but still carry latent information leakage risks.

**3. Diffusion Models**

Diffusion models generate data by gradually transforming noise into structured samples.

- High-quality outputs
- Improved robustness
- Increasingly used in modern generative AI

However, they may still capture sensitive patterns from training data.

**4. Statistical and Rule-Based Methods**

**Traditional approaches include:**

- Bayesian networks
- Copula-based models

- Sampling techniques

**These methods are:**

- More interpretable
- Less prone to memorization
- But often less expressive than deep learning models

**Applications of Synthetic Data**

Synthetic data is widely used for:

- Healthcare: Patient data sharing without violating privacy laws
- Finance: Fraud detection and risk modeling
- Autonomous Systems: Training perception models
- Cybersecurity: Simulating attack scenarios

It enables safe data sharing, model training, and testing environments.

**Advantages of Synthetic Data**

- Preserves privacy (to an extent)
- Enables data sharing across institutions
- Reduces regulatory constraints
- Supports data augmentation for machine learning

**Limitations and Security Concerns**

**Despite its benefits, synthetic data introduces several risks:**

- Memorization of training data
- Membership inference vulnerability
- Model inversion and reconstruction attacks
- Bias amplification from original datasets

Generative models may inadvertently encode sensitive attributes, leading to model leakage, especially when trained on small or highly unique datasets.

### 2.2 Membership Inference Attacks (MIAs)

Membership inference attacks attempt to answer:
"Was this data point used in training the model?"

**Attack types:**

- Black-box attacks (query access only)
- White-box attacks (full model access)
- Shadow model attacks

MIAs exploit differences in model behavior between training and non-training samples.

### 2.3 Model Leakage

Model leakage refers to unintended exposure of training data information through:

- Memorization
- Overfitting
- Latent space reconstruction

Generative models are particularly susceptible due to their objective of replicating data distributions.

## 3. Threat Model

| Component | Description |
|---|---|
| Target Model | Generative model trained on sensitive data |
| Adversary Knowledge | Partial or full knowledge of data distribution |
| Access Type | Black-box or white-box |
| Goal | Identify membership or reconstruct training samples |

## 4. Attack Mechanisms

4.1 Membership Inference via Representation Learning

Recent work proposes using contrastive representation learning to detect similarities between synthetic and real records. (PMC)

**Steps:**

1. Extract latent representations
2. Compute similarity scores
3. Infer membership based on distance metrics

### 4.2 Overfitting-Based Attacks

Overfitted models tend to:

- Assign lower loss to training samples
- Generate outputs similar to real records

Density-based methods (e.g., DOMIAS) exploit this behavior.

### 4.3 Distributional Leakage Attacks

**Adversaries analyze:**

- Clusters in synthetic data
- High-density regions representing real samples

These clusters act as proxies for original training data.

## 5. Vulnerability Analysis

Table 1: Vulnerability Comparison Across Synthetic Data Types

| Synthetic Data Type | Privacy Risk | Utility | Vulnerability to MIA |
|---|---|---|---|
| Fully Synthetic | Low | Moderate | Low |
| Partially Synthetic | High | High | High |
| Differentially Private Synthetic | Very Low | Reduced | Very Low |
| GAN-based Synthetic | Medium | High | Medium-High |
| VAE-based Synthetic | Medium | Moderate | Medium |

## 6. Experimental Findings (Literature-Based)

### 6.1 Key Observations

- Partially synthetic datasets show high susceptibility to membership inference attacks (ScienceDirect)
- Fully synthetic datasets are less vulnerable, but not completely secure (PMC)
- Overfitting significantly increases leakage risk
- Differential privacy reduces leakage but impacts data utility

### 6.2 Performance Metrics

| Metric | Description |
|---|---|
| Attack Accuracy | Probability of correct membership detection |
| Precision | Correct positive predictions |
| Recall | Detection rate of true members |
| Privacy Gain | Reduction in leakage risk |

### 6.3 Example Results (From Studies)

| Model Type | Attack Accuracy | Privacy Risk |
|---|---|---|
| GAN (No DP) | 70–90% | High |
| VAE-GAN Hybrid | 40–60% | Moderate |
| DP-GAN | 20–30% | Low |

Synthetic data generated without privacy constraints can still leak sensitive information due to memorization effects (ScienceDirect).

## 7. Model Leakage Risks

### 7.1 Memorization Effects

- Generative models may:
- Reproduce rare or unique records
- Leak identifiable patterns

### 7.2 Reconstruction Attacks

- Attackers reconstruct:
- Sensitive attributes
- Approximate training samples

### 7.3 Latent Space Exploitation

- Latent vectors may encode:
- Individual-level features
- Hidden correlations

## 8. Mitigation Strategies

### 8.1 Differential Privacy (DP)

- Adds noise during training
- Provides formal privacy guarantees
- Trade-off: reduced utility

### 8.2 Regularization Techniques

- Dropout
- Early stopping
- Weight decay

### 8.3 Model Auditing

- Evaluate MIA vulnerability before deployment
- Use privacy risk assessment frameworks

### 8.4 Synthetic Data Evaluation Metrics

| Metric | Purpose |
|---|---|
| Fidelity | Data realism |
| Diversity | Avoid memorization |
| Privacy Score | Leakage risk |

## 9. Discussion

The assumption that synthetic data is inherently privacy-preserving is flawed. While it reduces direct exposure, indirect leakage through models remains a critical concern. The trade-off between data utility and privacy is central to synthetic data research.

- Emerging challenges include:
- Evaluating privacy in large-scale generative models
- Handling bias amplification
- Ensuring robustness against adaptive attackers

**10. Conclusion**

This study highlights that synthetic data generation, while promising, introduces new security risks. Membership inference attacks and model leakage demonstrate that generative models can expose sensitive training information. Fully synthetic and differentially private methods provide stronger protection, but no approach is completely risk-free. Future research should focus on developing robust privacy guarantees, standardized evaluation metrics, and secure generative frameworks.

**References**

1. Zhang, Z., Yan, C., & Malin, B. (2022). Membership inference attacks against synthetic health data. Journal of Biomedical Informatics, 125, 103977. (PMC)
2. Hyeong, J., Kim, J., Park, N., & Jajodia, S. (2022). Membership inference attacks on tabular data synthesis models. arXiv.
3. Breugel, B. V., et al. (2023). Membership inference attacks via overfitting detection. arXiv.
4. Mustaqim, S. M., et al. (2025). Hidden data leakage in generative AI models. arXiv.
5. Liu, X., et al. (2024). Synthetic data for enhanced privacy: A VAE-GAN approach. Knowledge-Based Systems. (ScienceDirect)
6. Shokri, R., et al. (2017). Membership inference attacks against machine learning models. IEEE S&P.
7. Dwork, C. (2006). Differential privacy. ICALP.