



# From Real to Synthetic: GAN and DPGAN for Privacy Preserving Classifications

Mohammad Emadi<sup>1</sup>, Vahideh Moghtadaiee<sup>1</sup> <sup>a</sup> and Mina Alishahi<sup>2</sup> <sup>b</sup>

<sup>1</sup>*Cyberspace Research Institute, Shahid Beheshti University, Tehran, Iran*

<sup>2</sup>*Department of Computer Science, Open Universiteit, The Netherlands*

**Keywords:** Data Privacy, GAN, DPGAN, Synthetic Data, Classifier.

**Abstract:** Generative Adversarial Networks (GANs) and Differentially Private GANs (DPGANs) have emerged as powerful tools for generating synthetic datasets while preserving privacy. In this work, we investigate the impact of using GAN- and DPGAN-generated datasets on the performance of machine learning classifiers. We generate synthetic datasets using both models and train a variety of classifiers to evaluate their accuracy and robustness on multiple benchmark datasets. We compare classifier performance on real versus synthetic datasets in four different evaluation scenarios. Our results provide insights into the feasibility of using GANs and DPGANs for privacy-preserving data generation and their implications for machine learning tasks.

## 1 INTRODUCTION


Advancements in hardware and software have led to an explosion of data across various domains, fueling the progress of machine learning. However, many real-world applications, particularly in healthcare, face significant challenges in accessing sufficient training data (Ghosheh et al., 2024). Medical datasets are inherently limited due to the uniqueness of patient cases, the complexity of medical conditions, and ethical constraints on data sharing. Furthermore, strict privacy regulations and concerns about data confidentiality make large-scale data collection and distribution impractical (Xie et al., 2018). These limitations hinder the development and deployment of robust machine learning models in critical fields where high-quality data is essential.


To address this challenge, Generative Adversarial Networks (GANs) have emerged as a powerful tool for producing synthetic datasets that closely resemble real data while mitigating privacy risks (Goodfellow et al., 2014). However, standard GANs can still inadvertently expose sensitive information. Differentially Private GANs (DPGANs) (Xie et al., 2018) enhance privacy protection by integrating differential privacy mechanisms into the GAN framework, ensuring stronger guarantees against data

leakage. By incorporating noise during training, DPGANs provide provable privacy assurances while generating high-fidelity synthetic data.

This research investigates the trade-off between privacy preservation and classification performance when using Generative Adversarial Networks (GANs) and Differentially Private GANs (DPGANs) for synthetic data generation. We evaluate the effectiveness of these techniques by analyzing the accuracy of classification models trained on both real and synthetic datasets. Our study considers four experimental scenarios: training classifiers on real and synthetic data while testing them on both real and synthetic datasets. We conduct experiments on six benchmark datasets using five widely adopted classification algorithms, employing multiple performance evaluation metrics. Additionally, we examine the impact of varying the privacy budget on model performance, and on privacy gain using entropy metric.

Our results reveal that, despite a minor reduction in accuracy, synthetic data produced by DPGANs can successfully preserve privacy without significantly diminishing the performance of classification models. This study highlights the significant potential of DPGANs in enhancing data analysis while preserving privacy, offering valuable insights into the intersection of data privacy and machine learning.

<sup>a</sup>  <https://orcid.org/0000-0001-9655-4451>

<sup>b</sup>  <https://orcid.org/0000-0002-1159-8832>

## 2 RELATED WORK

In recent years, privacy-preserving machine learning has attracted significant attention. Techniques such as encryption, anonymization, (local) DP, and GANs have proven effective for safeguarding user privacy. Among these, GANs have been applied in various real-world scenarios (Ghosheh et al., 2024). For example, Moghtadaiee et al. (Moghtadaiee et al., 2025) propose differentially private GANs to generate synthetic indoor location data, ensuring data utility while protecting privacy through noise addition. Similarly, Zhang et al. (Zhang et al., 2021) integrate DPGANs within federated learning to detect COVID-19 pneumonia, enabling hospitals to collaboratively train models without sharing raw data.

Few studies have compared classifier performance under privacy-preserving techniques (Alishahi and Moghtadaiee, 2023). Research in (Alishahi and Zannone, 2021), (Lopuhaä-Zwakenberg et al., 2021), and (Sheikhalishahi and Zannone, 2020) investigates the effects of anonymization, DP, and encryption on classifiers' performance. The use of GANs for generating synthetic data to train classifiers has been briefly explored. Dat et al. (Dat et al., 2019) show that classifiers trained on GAN-generated data achieve satisfactory performance. Rashid et al. (Rashid et al., 2019) demonstrate that synthetic data can significantly improve skin lesion classification accuracy. However, the effect of DPGAN-generated data on classifier performance, and its comparison to GAN-generated data, remains largely unexplored. This work aims to address that gap.

## 3 SYSTEM MODEL

The architecture of the suggested system model is depicted in Fig. 1. It involves a comprehensive process for generating synthetic data using DPGAN, and subsequently evaluating its utility with various classifiers. Each section of the figure, labeled from (a) to (f), corresponds to a specific part of this process explained below:

**(a) DPGAN Training:** The process starts with the generator which creates synthetic data from a latent space. This latent space is a lower-dimensional space where random vectors are sampled and transformed into synthetic data samples that aim to mimic real data distribution. The discriminator then receives both real and synthetic data and tries to distinguish between them. It assigns a probability score to indicate whether the data is real or fake. To ensure DP, noise is added to the gradients during

the training of the discriminator. This process helps to prevent overfitting to the real data and ensures that the model does not memorize specific details of the training data, thus providing privacy guarantees. Loss for both networks is calculated based on their performance in distinguishing real from synthetic data. The generator aims to minimize this loss by producing more realistic data, while the discriminator tries to maximize that. This adversarial process is conducted by the backpropagation algorithm, which updates the weights of both networks based on the computed gradients to iteratively enhance the generator's synthetic data quality and the discriminator's discriminative power.

**(b) Generating Synthetic Dataset:** In this part, the trained generator is used to generate high-quality synthetic data.

**(c) Scenarios:** Four scenarios are considered for evaluating the generated synthetic data:

- **Scenario 1: Training and testing classifiers on real data (Real-Real):** This scenario serves as the baseline to determine the highest achievable performance when no synthetic data is involved.
- **Scenario 2: Training and testing on synthetic data (Fake-Fake):** This scenario evaluates classifiers trained and tested on DPGAN-generated synthetic data.
- **Scenario 3: Training on real data and testing on synthetic data (Real-Fake):** This scenario assesses how well classifiers trained on real data generalize to synthetic data.
- **Scenario 4: Training on synthetic data and testing on real data (Fake-Real):** This scenario evaluates whether classifiers trained solely on synthetic data can generalize to unseen real data, highlighting the practical utility of synthetic data in real-world applications.

**(d) Classifiers:** The synthetic and real data based on the four scenarios are evaluated using k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Logistic Regression (LR), Decision Tree (DT), and Random Forest (RF).

**(e) Evaluation Metrics:** The performance of classifiers is assessed by evaluation metrics such as Accuracy, Precision, Recall, and F1-Score. In section 4, we explain these four metrics.

**(f) Data Analysis:** The final part involves the analysis of results to ensure the quality and effectiveness of the synthetic data for various data analysis tasks.

The system model in Fig. 1 highlights the flow from data generation to performance evaluation, showing how the synthetic data by DPGAN is utilized and analyzed to ensure its quality and effectiveness

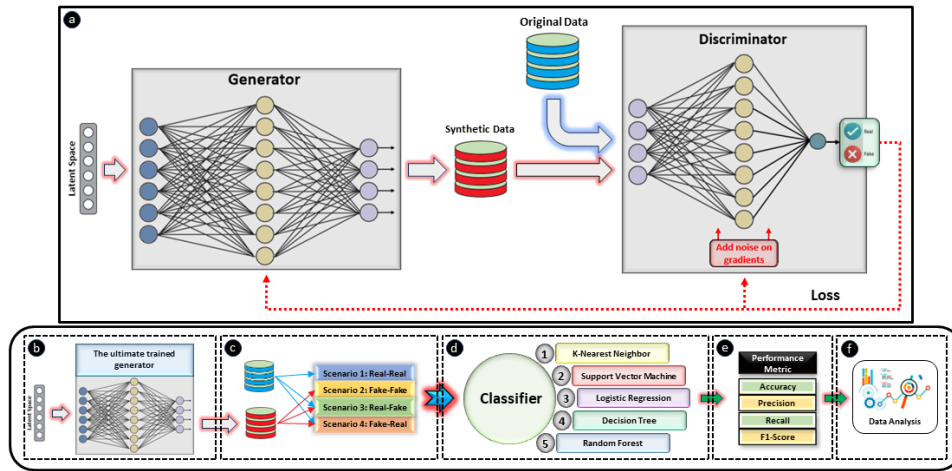


Figure 1: The architecture of the suggested system model.

Table 1: Summary of the datasets used in our study.

Dataset	Size	Features	Class Labels
Adult	48,842	14	$\leq 50K, > 50K$
Credit	690	15	+ (approved), - (rejected)
Mushroom	8,124	22	$e$ (edible), $p$ (poisonous)
Heart	299	12	0 (no event), 1 (heart failure)
Bankruptcy	10,503	64	0 (non-bankruptcy), 1 (bankruptcy)
Diabetic	1,151	19	0 (absence), 1 (presence)

for various data analysis tasks. The model emphasizes balancing data utility and privacy through the addition of noise during the training of the DPGAN, showing how the synthetic data can be suitable for practical applications while maintaining privacy.

## 4 EXPERIMENTAL SETUP

This section presents the experimental setup of our study. We conducted experiments using six different datasets and evaluated the performance of various classification models. The accuracy of both synthetic and real data was assessed under different scenarios, including training and testing with real data, synthetic data and combinations of both. Additionally, we explore the privacy-preserving capabilities of synthetic data generated using DPGAN varying privacy budget values. We utilize six datasets here, described below with its size, number of features, and class labels as reported in Table 1<sup>1</sup>.

### 4.1 Evaluation Metrics

This section presents evaluation metrics for assessing model performance. The data is split into training and testing sets, with the model trained on the former and evaluated on the latter using accuracy,

precision, recall, and F1-score. Additionally, two privacy evaluation metrics are employed: the privacy budget ( $\epsilon$ ) and entropy. In the following, we provide detailed explanations for each metric and their role in assessing the privacy levels of synthetic data.

**Privacy Budget ( $\epsilon$ ):** The privacy budget, denoted as  $\epsilon$ , is a key parameter in DP (Dwork et al., 2006) that balances privacy and data utility. In DPGANs, it controls the noise added to the discriminator's gradients during training to protect individual contributions. Lower  $\epsilon$  values imply stronger privacy through higher noise levels, while higher values reduce noise, improving utility but risking privacy leakage. In this study, we evaluate different  $\epsilon$  values to assess the impact of DP levels on synthetic data utility.

**Entropy:** Entropy (Kim et al., 2016), a concept from information theory introduced by Claude Shannon, measures the uncertainty or randomness within a probability distribution. For a discrete random variable  $X$  with possible outcomes  $x_1, x_2, \dots, x_n$  and associated probabilities  $P(x_1), P(x_2), \dots, P(x_n)$ , the entropy  $H(X)$  is:

$$H(X) = - \sum_{i=1}^n P(x_i) \log(P(x_i)) \quad (1)$$

where  $P(x_i)$  is the probability of outcome  $x_i$ . The logarithmic base chosen in entropy calculations determines its units. This formula calculates the expected amount of "information" or "surprise" in the variable  $X$ . Higher entropy values indicate greater uncertainty in predictions, meaning the data distribution is more randomized. In synthetic data, high entropy reflects greater diversity, enhancing privacy by reducing the risk of identifying specific data points. In this study, entropy is used to evaluate privacy protection in real and synthetic data.

<sup>1</sup><https://archive.ics.uci.edu/datasets>

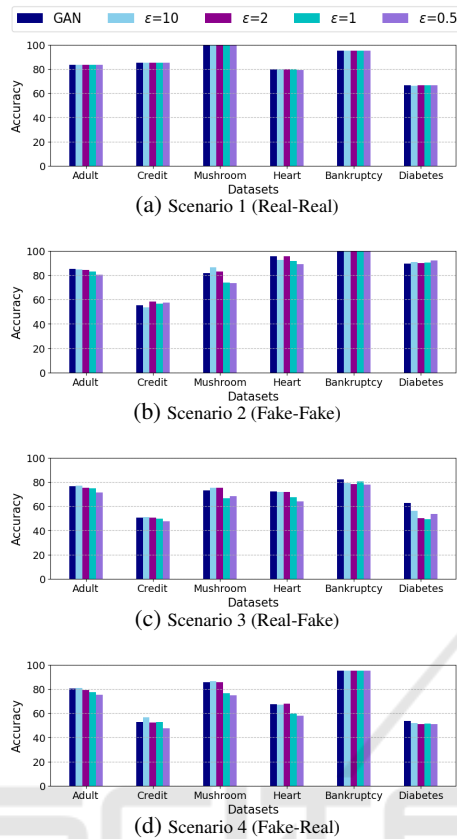


Figure 2: Average accuracy of five classifiers in different scenarios for GAN and DPGAN varying privacy budgets.

It is calculated for each feature, with the average representing overall dataset entropy. Comparing real and synthetic data entropy helps assess how closely synthetic data resembles real data. Higher synthetic data entropy suggests better privacy preservation by increasing diversity and reducing the likelihood of identifying specific information.

## 5 EXPERIMENTAL RESULTS

This section provide the utility and privacy results applying the proposed system model to our datasets.

### 5.1 Utility Evaluation Results

The accuracy of the synthetic data is evaluated using KNN, SVM, LR, DT, and RF across all four scenarios. Synthetic data is generated using DPGAN with privacy budgets (0.5, 1, 2, 10) to analyze the trade-offs between privacy and data utility. Figure 2 presents the classification accuracy for Real-Real, Fake-Fake, Real-Fake, and Fake-Real scenarios using synthetic data from regular GAN and DPGAN

with varying privacy budgets. To assess scenario influence, we computed the average accuracy across all classifiers.

The Bankruptcy dataset achieves the highest accuracy across all four scenarios, followed by the Mushroom and Adult datasets with nearly equal accuracies, and then the Heart dataset. The Credit and Diabetes datasets show the lowest and nearly equal accuracies. Accuracy decreases with lower privacy budgets across all scenarios. However, even with a low privacy budget of 0.5, the accuracy for each dataset remains stable compared to the case without differential privacy (using GAN) and stays within an acceptable range.

We can infer that the type of features affects model accuracy; datasets with more integer or continuous features, like the Bankruptcy dataset, achieve higher accuracies. Conversely, the size of the dataset also influences accuracy, with larger datasets, such as the Adult dataset, showing high accuracy despite having categorical features. Smaller datasets, like Credit and Diabetes, exhibit lower accuracies. Therefore, both feature type and dataset size are key factors in determining classification model performance.

In addition, in Scenario 1 (Real-Real), where both training and testing use real data, accuracy is highest and serves as a baseline. In Scenario 2 (Fake-Fake), where synthetic data is used for both training and testing, GAN and DPGAN maintain high accuracy across varying privacy budgets, especially for the Bankruptcy and Mushroom datasets. In Scenario 3 (Real-Fake), where classifiers are trained on real data and tested on synthetic data, accuracy slightly drops, notably for the Credit and Diabetes datasets, likely due to class imbalance affecting GAN's generation quality. Similarly, in Scenario 4 (Fake-Real), where classifiers are trained on synthetic data and tested on real data, accuracy is comparable to Scenario 3. Overall, while lower privacy budgets cause a slight decline, synthetic data remains effective for training classification models.

Note that the results of Scenario 4 are particularly significant as they demonstrate the potential to train classification models exclusively with synthetic data, eliminating the need for real data to address privacy concerns. This finding highlights the ability of models to generalize and accurately predict outcomes for new, unseen real data based solely on synthetic data. This capability ensures data privacy while demonstrating the robustness and reliability of the synthetic data. Such an approach has the potential to transform data privacy practices by enabling the development of effective machine learning models without compromising sensitive information.



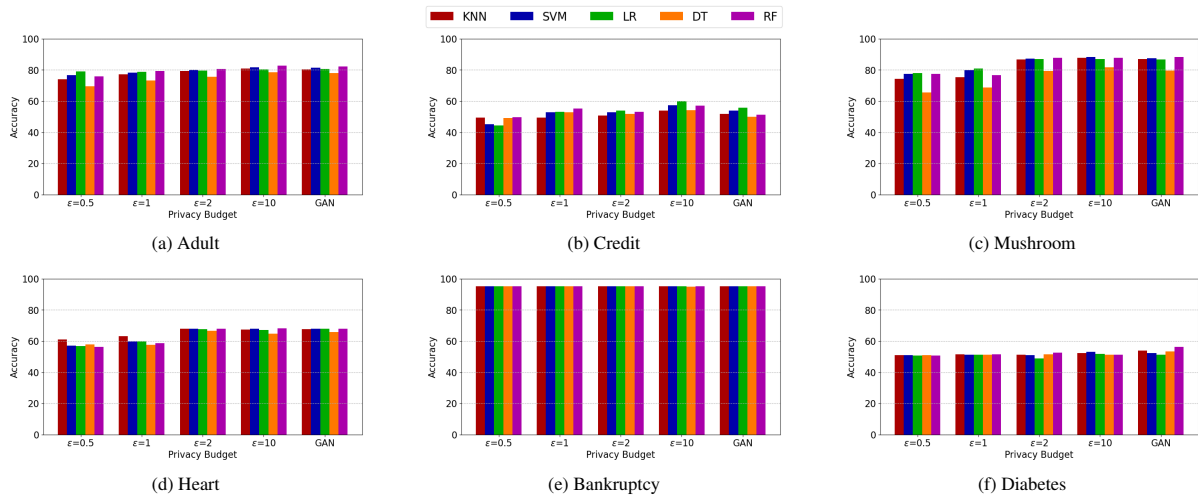


Figure 3: Accuracy for Scenario 4 across different datasets and classifiers, showing the impact of varying privacy budgets.

Table 2: Average performance over five classifiers for GAN and DPGAN for different  $\epsilon$  values in Scenario 4 (Fake-Real).

Dataset	Metric	Privacy Budget				
		$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 10$	GAN
Adult	Accuracy (%)	75.02	77.35	79.00	80.82	80.56
	Precision (%)	76.04	76.50	78.08	79.98	79.44
	Recall (%)	75.02	77.35	79.00	80.82	80.56
	F1-Score (%)	73.69	75.92	77.74	79.59	79.30
Credit	Accuracy (%)	47.48	52.67	52.44	56.53	52.50
	Precision (%)	47.57	51.32	50.92	56.65	51.89
	Recall (%)	47.48	52.67	52.44	56.53	52.50
	F1-Score (%)	42.19	47.85	48.17	54.76	50.91
Mushroom	Accuracy (%)	74.56	76.34	85.60	86.52	85.81
	Precision (%)	75.28	77.15	86.32	87.37	86.38
	Recall (%)	74.56	76.34	85.60	86.52	85.81
	F1-Score (%)	74.23	76.20	85.50	86.41	85.74
Heart	Accuracy (%)	57.84	59.79	67.64	67.02	67.45
	Precision (%)	52.27	53.17	51.05	52.22	50.99
	Recall (%)	57.84	59.79	67.64	67.02	67.45
	F1-Score (%)	52.36	52.24	55.83	57.01	55.76
Bankruptcy	Accuracy (%)	95.24	95.22	95.21	95.16	95.23
	Precision (%)	90.87	90.99	90.94	90.95	90.92
	Recall (%)	95.24	95.22	95.21	95.16	95.23
	F1-Score (%)	92.97	92.97	92.96	92.94	92.97
Diabetes	Accuracy (%)	50.83	51.36	50.99	51.92	53.36
	Precision (%)	46.95	47.34	48.91	50.80	53.10
	Recall (%)	50.83	51.36	50.99	51.92	53.36
	F1-Score (%)	42.06	46.61	47.21	48.59	51.88

Table 2 presents the average results for Accuracy, Precision, Recall, and F1-Score in Scenario 4 (Fake-Real) across all datasets. The results show a decreasing trend in all metrics as privacy increases (i.e., privacy budget decreases), highlighting the trade-off between model performance and privacy protection. Lower privacy budgets reduce available information, leading to decreased accuracy. However, some datasets maintain acceptable performance even at lower privacy budgets, suggesting privacy-preserving techniques can be applied effectively without major accuracy loss.

While Table 2 summarizes average results (Fake-Real scenario), Fig. 3 focuses on accuracy per dataset, illustrating variations with different privacy budgets. Accuracy generally improves with higher privacy budgets across datasets. Notably, RF

and LR models achieve higher accuracy, with RF showing greater stability due to result aggregation across multiple trees, whereas DT is less stable under varying privacy conditions. These findings highlight that both model choice and privacy budget significantly influence classification performance, with RF and LR proving most effective across privacy scenarios.

## 5.2 Privacy Evaluation Results

To evaluate the privacy of synthetic data from GAN and DPGAN, we assess entropy as a measure of data diversity and privacy preservation across different privacy budgets. Table 3 compares entropy values of real and synthetic data for each dataset, covering privacy budgets from  $\epsilon = 0.5$  to  $\epsilon = 10$  and a non-private GAN. The results show that synthetic data consistently exhibits higher entropy than real data, indicating greater uncertainty and variability. This increased entropy makes re-identification harder, demonstrating that both GAN and DPGAN enhance privacy by generating more diverse data.

A closer analysis shows that the Bankruptcy dataset has the highest increase in entropy, while the Mushroom dataset shows the least. This discrepancy stems from the data nature; entropy captures variations more effectively in numerical datasets like Bankruptcy, better reflecting privacy improvements. In contrast, Mushroom, being entirely categorical, is less impacted by entropy measures. Among the numerical datasets (Bankruptcy, Diabetes, and Heart), Bankruptcy's greater entropy increase is likely due to its larger volume and higher diversity, allowing the model to generate more complex and varied synthetic data, enhancing privacy preservation.

Table 3: This table compares the entropy values of real and synthetic datasets across different datasets. It includes the entropy values of the real data alongside the entropy values of the synthetic data under various privacy budget parameters.

Dataset	Real Entropy	Fake Entropy					Fake Entropy
		$\epsilon = 0.5$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 10$	GAN	Average
Adult	9.2969	10.3115	10.3636	10.3742	10.2718	10.3306	10.3303
Credit	4.6713	6.0720	6.0913	6.0595	6.0561	6.0619	6.0682
Mushroom	8.5522	8.5704	8.5642	8.5737	8.5565	8.5631	8.5656
Heart	4.5593	5.1202	5.1318	5.1892	5.1805	5.1609	5.1565
Bankruptcy	1.4395	8.7104	8.6884	8.6949	8.6709	8.2461	8.6021
Diabetes	4.5782	6.4472	6.4791	6.5277	6.5613	6.5278	6.5086

Additionally, these differences can be attributed to the datasets' nature, complexity, and feature diversity. Simpler datasets like Mushroom, with limited attributes and predictable distributions, make it harder for GAN and DPGAN models to introduce complexity, resulting in minimal entropy changes. In contrast, complex datasets like Bankruptcy, with heterogeneous features and less predictable patterns, drive GAN models to generate more diverse synthetic data, leading to higher entropy and better privacy gain. Notably, Bankruptcy is the only dataset showing a clear increase in entropy as  $\epsilon$  decreases, likely due to its structural complexity enabling greater diversity under tighter privacy constraints. However, this does not mean similar effects are absent in other datasets, as entropy alone may not fully reflect privacy improvements, especially in simpler or categorical datasets.

## 6 CONCLUSIONS

In this paper, we explored the use of GANs and Differentially Private GANs (DPGANs) based on the Wasserstein distance to generate synthetic datasets that balance utility and privacy for classifier training. Comparing classifiers trained on real and synthetic data, we found that although there is a slight accuracy trade-off, synthetic data from GAN and DPGAN effectively preserves privacy without major performance loss. Additionally, higher entropy levels in synthetic data reflect greater randomness and diversity, making it harder to link synthetic samples to real data and enhancing privacy protection.

## REFERENCES

Alishahi, M. and Moghtadaiee, V. (2023). *Collaborative Private Classifiers Construction*, pages 15–45. Springer, Cham.

Alishahi, M. and Zannone, N. (2021). Not a free lunch, but a cheap one: On classifiers performance on anonymized datasets. In *Data and Applications Security and Privacy*, volume 12840, pages 237–258. Springer.

Dat, P. T., Dutt, A., Pellerin, D., and Quénot, G. (2019). Classifier training from a generative model. In *International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE.

Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. (2006). Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology—EUROCRYPT 2006*, pages 486–503.

Ghosheh, G. O., Li, J., and Zhu, T. (2024). A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Comput. Surv.*, 56(6).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Kim, S.-H., Jung, C., and Lee, Y.-J. (2016). An entropy-based analytic model for the privacy-preserving in open data. In *International Conference on Big Data*, pages 3676–3684. IEEE.

Lopuhaä-Zwakenberg, M., Alishahi, M., Kivits, J., Klarenbeek, J., van der Velde, G., and Zannone, N. (2021). Comparing classifiers' performance under differential privacy. In *International Conference on Security and Cryptography, SECRYPT*, pages 50–61.

Moghtadaiee, V., Alishahi, M., and Rabiei, M. (2025). Differentially private gans for generating synthetic indoor location data. *International Journal of Information Security*, 24:1–21.

Rashid, H., Tanveer, M. A., and Khan, H. A. (2019). Skin lesion classification using gan based data augmentation. In *International conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 916–919. IEEE.

Sheikhalishahi, M. and Zannone, N. (2020). On the comparison of classifiers' construction over private inputs. In *International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 691–698.

Xie, L., Lin, K., Wang, S., Wang, F., and Zhou, J. (2018). Differentially private generative adversarial network.

Zhang, L., Shen, B., Barnawi, A., Xi, S., Kumar, N., and Wu, Y. (2021). Feddpagan: federated differentially private generative adversarial networks framework for the detection of covid-19 pneumonia. *Information Systems Frontiers*, 23(6):1403–1415.