P. Sanchez-Serrano and R. Rios and I. Agudo, "Privacy-preserving tabular data generation: Systematic Literature Review", *19th DPM International Workshop on Data Privacy Management (DPM 2024)*, In press.
NICS Lab. Publications: https://www.nics.uma.es/publications

Privacy-preserving tabular data generation: Systematic Literature Review

Pablo Sanchez-Serrano, Ruben Rios, and Isaac Agudo

Network, Information and Computer Security (NICS) Lab, University of Malaga, Spain {pablosanserr,ruben.rdp,isaac}@uma.es

Abstract. There is a wide range of tabular data of great value to science, economy and social progress. When sharing such data, privacy must be taken into account. Traditionally, this has been addressed through anonymization. However, in recent years, with the growth of AI, the possibility of using generative models has emerged as a way to generate synthetic data that guarantees privacy while maintaining their utility. This systematic literature review aims to identify and classify existing privacy-preserving tabular generative models in order to create a taxonomy of solutions. In addition, we analyze the privacy metrics and techniques they use, and identify possible unexplored lines of research.

Keywords: Synthetic data · Privacy · Tabular data.

1 Introduction

There is a wide variety of tabular data, including medical records, financial transactions, and demographic details. This data holds immense value for scientific, economic and social progress, as it can be used to identify patterns, facilitate decision-making and disseminate knowledge. However, the sharing of this data raises privacy concerns, given that it often contains PII (personally identifiable information).

Traditional methods for protecting privacy in tabular data include [20]: data pseudonymization, which replaces PII with fake identifiers, and data anonymization, which involves generalization, suppression and perturbation techniques that modify attributes in the dataset to obtain a supposedly anonymous dataset. To decrease the risk of re-identification some models like k-anonymity, l-diversity and t-closeness have been proposed. Recently, generative models have emerged as a way to guarantee the privacy of datasets [9]. These models generate synthetic data from real datasets, mimicking the statistical properties of the training data.

When dealing with synthetic datasets, there are significant differences in the amount of knowledge and access available to different users (see Fig. 1). This involves a range of privacy challenges that need to be considered. Users further to the right of the diagram show a higher level of difficulty in discerning which data were used to generate the synthetic data. The number of barriers will be higher the further to the right the user is located, i.e. the less knowledge and access the user has.

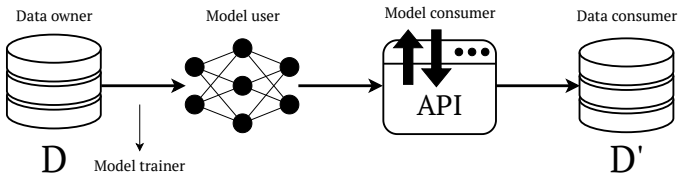

Fig. 1. Different levels of knowledge and access to the trained model.

A *model trainer* uses the real data (D) given by the data owner to train a generative model. The model trainer must be careful with possible data leakage due to errors or intermediate outputs. The model trainer could also be malicious, or the data owners may not trust the data owner. Security mechanisms such as homomorphic encryption [3] or federated learning [36] should be implemented. Once the model is trained, the user can have different levels of access to the model. We refer to the user with full access to the model as the *model user*. Despite having completed the training phase, it may be possible to obtain information about D from the model [38]. Conversely, a *model consumer* can only generate samples from the model using an API, but do not have access to the trained model. The amount of information available to this type of users depends on the API. A first-level API allows unlimited samples generation, leading to honest-but-curious users who seeks information while respecting established protocols. On the other hand, a second-level API has some restrictions on data generation, i.e. limited number of requests or attributes that are not allowed to be generated. Membership Inference Attacks (MIAs) [26] can exploit the lack of restrictions on data generation. MIAs take advantage of differences in how models respond to queries from members inside and outside of the training dataset. Finally, the *data consumer* only has access to a synthetic dataset (D') generated by the model, and is unable to generate samples by himself. Although more challenging, it is possible to obtain information about D from D' [4].

The contributions of this paper can be summarized as follows:

1. The use of a systematic methodology to provide an overview of privacy techniques used in tabular data generative models.
2. A collection of 24 systematically selected papers.
3. A collection of privacy metrics for in tabular data generative models.
4. A taxonomy of privacy-preserving generative models for tabular data.

This works is organized as follows. Section 2 introduces the methodology and how the papers were selected. Section 3 discusses the different ways to measure privacy in tabular data generation and explains the techniques used to ensure privacy collected from the selected papers. Section 4 provides a taxonomy of generative models for tabular data, giving an order and clarifying the differences between them. Finally, Section 5 draws conclusions and outlines possible lines of future research based on the observations made in the paper.

2 Systematic literature review

A Systematic Literature Review (SLR) is a rigorous approach to reviewing and synthesizing research literature on a specific topic. This methodology is designed to provide a comprehensive, unbiased and reproducible summary of existing research. The PICOC framework is employed to define the scope and focus of our study. It involves three main steps: planning, conducting and reporting.

2.1 Planning

This SLR is performed to answer the following questions:

1. What are the main techniques used to guarantee privacy in generative models for tabular data?
2. How can we measure the privacy of generative models for tabular data?

PICOC terms help to define a list of keywords, as shown in Table 1. Using these keywords we can create a search query (see Definition 1), which addresses our research questions.

Table 1. Keyword list created from PICOC terms.

Keywords	Synonyms	PICOC
Tabular data	Database, Dataset	Population
Privacy techniques	Data masking, Differential privacy, Masked data, Privacy approach, Privacy methods, Privacy-preserving, k-anonymity, l-diversity, t-closeness	Intervention
Generative model	Data synthesis, Synthesizer, Synthetic data generation, Synthetic generator	Comparison
Benchmark		Outcome
Privacy metric	Anonymity metric	Outcome
Utility metric	Data quality, Data utility, ML efficacy, Usefulness of data	Outcome

Definition 1 (Search Query). *("Tabular data" OR "Database" OR "Dataset") AND ("Privacy techniques" OR "Data masking" OR "Differential privacy" OR "Masked data" OR "Privacy approach" OR "Privacy methods" OR "Privacy-preserving" OR "k-anonymity" OR "l-diversity" OR "t-closeness") AND ("Generative model" OR "Data synthesis" OR "Synthesizer" OR "Synthetic data generation" OR "Synthetic generator") AND ("Benchmark" OR "Privacy metric" OR "Anonymity metric" OR "Utility metric" OR "Data quality" OR "Data utility" OR "ML efficacy" OR "Usefulness of data")*

The next step is to define which digital libraries use to search. We selected IEEE Digital Library, ISI Web of Science and Scopus. There might be duplicate papers but this will be taken into account in the conducting phase.

To refine the search and ensure the inclusion of high-quality and relevant studies, the following exclusion criteria are applied: (i) accepted papers should address privacy for generative AI models for tabular data, (ii) surveys or reviews will be discarded, (iii) only articles, conference papers, proceedings or journals will be considered, (iv) a minimum number of citations is required. Papers published before 2022 should have at least 20 citations. Papers from 2022 are required to include a minimum of 10 citations. Papers from 2023 or 2024 must have a minimum of 5 citations. To sum up, these are the exclusion criteria:

- The paper does not discuss privacy
- The paper does not discuss AI
- The paper does not focus on tabular data
- It is a survey/review
- It is not an article, conference paper, proceeding or journal
- It has not enough citations
- It is not published in English

After an initial filtering using the exclusion criteria, a checklist of five questions (listed below) with specific criteria is established. There are three possible scores for each criterion: Yes (1 point), Partially (0.5 points), or No. Thus, 5 points is the maximum score. Papers that reach 3 points are finally selected.

1. Does the article propose a new AI model for tabular data generation?
2. Does the article propose new attacks to privacy in generative models?
3. Does the paper propose a model practical implementation?
4. Does the model include techniques to provide privacy?
5. Does the article discuss how to measure privacy for tabular generative data models? Does it also include a way to measure utility?

2.2 Conducting

The first step is to perform a search using the query string presented in Section 2.1. Initially, a total of 977 papers were found. From this list of papers, 36 were duplicated, giving a total of 941 unique papers. To provide a clearer understanding of the evolution of research on this topic, Figure 2 illustrates the number of papers published each year. The graph shows a growth in the number of papers over the years. Although the number of papers published in 2024 is lower than in previous years, the reason is that the current writing date is mid 2024.

This is the moment to apply the exclusion criteria presented in Section 2.1. All papers are reviewed, focusing on the title, keywords, and abstract. At the end of this process, 61 papers are accepted.

After an initial filtering, it is time to apply the Quality Assessment Checklist presented in Section 2.1. During this step, potential papers are added through snowballing. The papers added in this way are also submitted to Quality Assessment Checklist. During the conducting process, a backward snowballing (or backward reference searching) is performed. This involves looking through the references listed in the selected papers to find older studies that the key papers

Fig. 2. Number of papers found

Table 2. Reference list of papers.

Years	Papers
2017	[15]
2019	[2], [34], [11]
2020	[33], [12]
2021	[6], [27], [13], [5]
2022	[28], [31], [32], [30], [8], [16]
2023	[22], [14], [35] [17], [29], [18], [19]
2024	[37]

have cited, which might also be relevant in the research topic. At the end of this process, a final list of 24 papers are selected. The reference list of papers is shown in Table 2. As with the papers found with the query (Figure 2) there is an increase in the number of selected papers over the years, except in 2024.

2.3 Reporting

In this section, we extract some statistical data about the selected articles. The information extracted from the papers is discussed in the following sections.

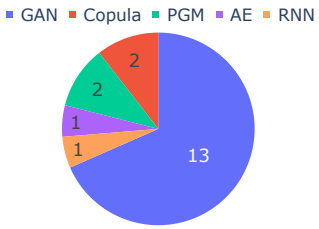

Fig. 3. Model family distribution, in which models are grouped according to their nature or type.

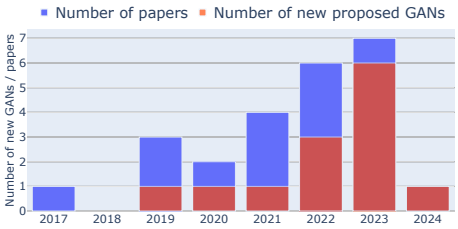

Fig. 4. Evolution of the GANs proposed in the selected papers compared to the number of selected papers.

Out of the 24 selected papers, 17 papers propose a new model for privacy-preserving tabular data generation. There are two papers that propose two models, for a total of 19 proposed models. Figure 3 shows the different types of model families collected. This chart will be useful in establishing a taxonomy of different generative models. There is a clear predominance of GANs over the others.

Figure 4 compares the years of creation of GANs with the years of publication of all selected papers. It can be seen that the growth of interest in GANs follows the growth of interest in the research area. This shows that GANs are the type of generative models that are most often used to generate tabular data with privacy guarantees.

3 Measuring Privacy in tabular data generation

There are several ways to measure privacy in generative models for tabular data. Some traditional privacy techniques, such as k-anonymity or t-closeness, can also implicitly act as privacy measures. Among the selected papers, differential privacy stands out.

Differential privacy [7] is a mathematical framework designed to provide privacy guarantees for data entries within a dataset. Differential privacy ensures that the inclusion or exclusion of a single individual's data does not significantly affect the outcome of any analysis, thereby protecting the individual's privacy.

Definition 2 (Neighboring Datasets). *Two datasets, D and D', are neighboring, if and only if D' differs from D in only one entry.*

Definition 3 ((ε, δ)-Differential Privacy). *For a non-negative privacy budget ε and a non-negative relaxation term δ, an algorithm, M, satisfies (ε, δ)-differential privacy if for any pair of neighboring datasets D, D' and $S \subseteq Range(M)$*

$$Pr[M(D) \in S] \leq exp(\varepsilon) \cdot Pr[M(D') \in S] + \delta \tag{1}$$

where Pr is taken with respect to the randomness of M. δ is a relaxation term to ε-differential privacy. There are a variety of techniques for achieving differential privacy. Essentially, the algorithm M perturbs the input with some noise distribution, i.e. normal distribution, based on ε and δ.

The following expression is obtained by clearing ε from expression 1:

$$\varepsilon \geq ln\left(\frac{Pr[M(D) \in S] - \delta}{Pr[M(D') \in S]}\right) \tag{2}$$

A lower value of ε implies a higher level of privacy because inequality 2 is more restrictive. However, decreasing ε increases the noise that needs to be added to satisfy Definition 3

There are some variations or extensions of the definition of differential privacy, such as RDP (Rényi Differential Privacy) [21], LDP (Local Differential Privacy) or CDP (Concentrated Differential Privacy).

Privacy accounting concept indicates that there is a need of some "accountant" procedure that computes the privacy cost at each access to the training data, and accumulates this cost as the training progress [1]. The privacy analysis of our some differential privacy techniques employs the moments accountant approach to keep track of the privacy cost in multiple iterations. This concept can also be used to measure privacy degradation with increasing number of queries. One way to compensate for this progressive loss of privacy would be to progressively increase the noise.

There are several techniques to ensure differential privacy, such as Differentially Private Expectation Maximization (DP-EM) [25], Private Aggregation of Teacher Ensembles (PATE) [23,24] or Differentially Private Stochastic Gradient

Descent (DP-SGD) [1]. In general, they all involve the addition of noise in one way or another.

Similar to differential privacy, there is also the concept of identifiability [33]. This framework is used to measure and limit the risk of re-identification. There are also other ways to measure privacy for those models that do not theoretically guarantee privacy, but rather focus on an empirical approach to measure privacy. These focus on performing attacks to see how effective they are. The most common is the Membership Inference Attack (MIA) [26].

SELENA [30] is a ensemble method that combines Split-AI and Self-Distillation to mitigate MIAs. Although SELENA is primarily designed for supervised classification tasks, it could be used as a component of a generative model. For example, SELENA could be used in GANs to protect the discriminator from revealing membership information about the training data. SELENA trains submodels on random data subsets and uses adaptive inference to ensure similar behavior on member and non-member inputs, significantly reducing MIA risks.

4 A taxonomy for tabular data generative models

This section categorizes tabular data generative models from selected papers (see Figure 5). Due to length restrictions, the taxonomy focuses on GANs with privacy guarantees. However, other types of models were found:

- Autoencoders (AEs): DP-SYN [2]
- Probabilistic Graphical Models (PGMs): PrivMRF [5] and PrivIncr [18]
- Recurrent Neural Networks (RNNs): Conditional-LSTM [22]
- Copula-based models: LoCop and DR_LoCop [32]

The white boxes in Figure 5 represent each of the 13 models, while the gray boxes represent the categories into which the different models fall. Note that DP-GAN, whose connector is shown as a dotted line, is a particular case. Although it is possible to introduce conditions on one of its components [13], it does not fall within the definition of a conditional GAN. Therefore, it is placed in the category of non-conditional GANs. Models that were originally designed to generate EHR (Electronic Health Record) data are in a green box. Similarly, those GANs that integrate an autoencoder as a component of their model are in a blue box.

4.1 Generative adversarial networks (GAN)

A generative adversarial network (GAN) [10] is a type of machine learning framework where two neural networks are trained simultaneously in a zero-sum game setting. GANs have established themselves as one of the state-of-the-art generative models. GANs consists of two adversarial models:

- Generator G: takes random noise as input and generates samples. It aims to generate data that imitates a given dataset.

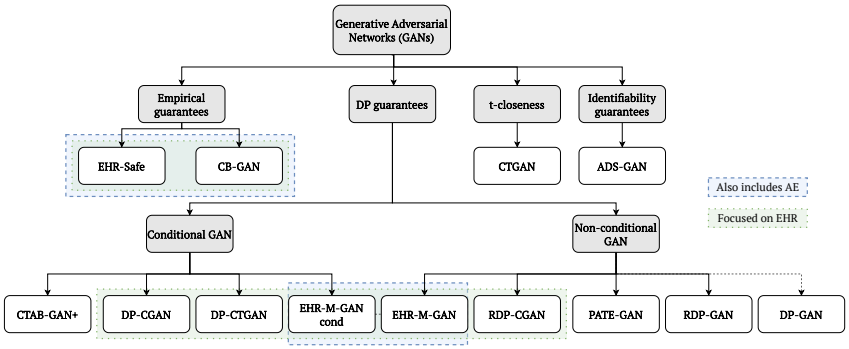

Fig. 5. Privacy-preserving tabular data GAN taxonomy

– Discriminator D: attempts to differentiate between real data samples taken from the training dataset and fake data samples generated by the generator. It outputs a probability indicating if a given sample is real or fake.

The generator tries to fool the discriminator by generating realistic data. The discriminator tries to become better at distinguishing real data from fake data. This creates a minimax game between them. The generator aims to maximize the probability of the discriminator misclassifying its outputs as real, and the discriminator aims to minimize the probability of incorrectly classifying real data as fake and vice versa.

There is a wide variety of GANs, each one specialized in generating certain kinds of data, such as images, video, network trafic, tabular data, etc.

Conditional GANs There is no control on the process of data generation in a standard GAN. It generates synthetic data from the real data without allowing any further conditions or requirements. Conditional Generative Adversarial Networks (CGANs) are used to address this problem. With CGANs, a condition can be included to control the data generation process. The following types of CGANs are designed to generate tabular data ensuring differential privacy:

– CTAB-GAN+ [37]: It is a general purpose model trained with DP-SGD to impose strict privacy guarantees and leverage the RDP for privacy accounting because it provides stricter bounds on the privacy budget.
– DP-CGAN [29]: It is focused on EHR data generation. This model uses standard differential privacy.
– DP-CTGAN [8]: It is focused on EHR data generation. This model also uses standard differential privacy. Has a federated learning-oriented variant, FDP-CTGAN.
– EHR-M-GAN cond [17]: It is a conditional variation of EHR-M-GAN. It is focused con EHR data generation. It uses a dual variational autoencoder (dual-VAE) as a part of its architecture. DP-SGD is used to guarantee privacy.

Non-conditional GANs There are other ways to create synthetic data with privacy assurances beyond CGANs. The following GAN models provide privacy guarantees but are not conditional:

- EHR-M-GAN [17]: It is focused con EHR data generation. It uses a dual variational autoencoder (dual-VAE) as a part of its architecture. It uses DP-SGD to guarantee privacy.
- DP-GAN [13]: One of the components is a conditional network, but it is not a conditional GAN as CGANs are defined. This model uses standard differential privacy.
- PATE-GAN [34]: This model modify the discriminator to be differentially private using a modified version of PATE framework.
- RDP-CGAN [31]: It is a convolutional GAN focused on EHR data. To ensure privacy, this model uses RDP.
- RDP-GAN [19]: This model uses RDP to ensure privacy. It is a general purpose model.

5 Conclusions

This paper provides an overview of the state of the art in privacy-preserving tabular data generation. From a total of 941 unique papers, we selected 24 papers to answer two research questions: "What are the main techniques used to guarantee privacy in generative models for tabular data?" and "How can we measure the privacy of generative models for tabular data?". For the first question, we found that although there is a wide range of generative models in the literature, GAN is the predominant model for synthetic tabular data generation, and the most used application scenario is the protection of medical records. Regarding the second question, most models focus on providing differential privacy guarantees, either its standard definition or some variants. However, we also found some models that do not theoretically guarantee privacy, but rather focus on an empirical approach to measure privacy. As future work, we plan to identify other generative models where the community has not yet begun to discuss privacy risks, and analyze the reasons for this, in order to incorporate privacy guarantees into these models.

Acknowledgments. This work has been partially supported by project PID2022-139268OB-I00, financed by MCIN/AEI /10.13039/501100011033 / FEDER, UE and project TED2021-129830B-I00, financed by MCIN/AEI /10.13039/501100011033/Next-GenerationEU/PRTR.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep Learning with Differential Privacy. In: Conference on Computer and Communications Security (CCS). p. 308–318. ACM (2016)

2. Abay, N.C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., Sweeney, L.: Privacy Preserving Synthetic Data Release Using Deep Learning. In: Machine Learning and Knowledge Discovery in Databases. pp. 510–526. Springer (2019)
3. Armknecht, F., Boyd, C., Carr, C., Gjøsteen, K., Jäschke, A., Reuter, C.A., Strand, M.: A Guide to Fully Homomorphic Encryption. Cryptology ePrint Archive, Paper 2015/1192 (2015)
4. van Breugel, B., Sun, H., Qian, Z., van der Schaar, M.: Membership inference attacks against synthetic data through overfitting detection (2023), `https://arxiv.org/abs/2302.12580`
5. Cai, K., Lei, X., Wei, J., Xiao, X.: Data synthesis via differentially private markov random fields. Proc. VLDB Endow. **14**(11), 2190–2202 (2021)
6. Domingo-Ferrer, J., Muralidhar, K., Bras-Amorós, M.: General Confidentiality and Utility Metrics for Privacy-Preserving Data Publishing Based on the Permutation Model. IEEE Trans on Dependable and Secure Computing **18**(5), 2506–2517 (2021)
7. Dwork, C.: Differential privacy. In: Automata, Languages and Programming. pp. 1–12. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
8. Fang, M.L., Dhami, D.S., Kersting, K.: DP-CTGAN: Differentially Private Medical Data Generation Using CTGANs. In: Artificial Intelligence in Medicine. pp. 178–188. Springer (2022)
9. Figueira, A., Vaz, B.: Survey on Synthetic Data Generation, Evaluation Methods and GANs. Mathematics **10**(15) (2022)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Commun. ACM **63**(11), 139–144 (2020)
11. Hittmeir, M., Ekelhart, A., Mayer, R.: Utility and Privacy Assessments of Synthetic Data for Regression Tasks. In: IEEE Conference on Big Data. pp. 5763–5772 (2019)
12. Hittmeir, M., Mayer, R., Ekelhart, A.: A Baseline for Attribute Disclosure Risk in Synthetic Data. In: ACM Conference on Data and Application Security and Privacy (CODASPY). p. 133–143. ACM (2020)
13. Ho, S., Qu, Y., Gu, B., Gao, L., Li, J., Xiang, Y.: DP-GAN: Differentially private consecutive data publishing using generative adversarial nets. Journal of Network and Computer Applications **185**, 103066 (2021)
14. Hu, R., Li, D., Ng, S.K., Zheng, Z.: CB-GAN: Generate Sensitive data with a Convolutional Bidirectional Generative Adversarial Networks. In: Database Systems for Advanced Applications. pp. 159–174. Springer Nature (2023)
15. Jia, R., Sangogboye, F.C., Hong, T., Spanos, C., Kjærgaard, M.B.: PAD: protecting anonymity in publishing building related datasets. In: ACM Conference on Systems for Energy-Efficient Built Environments (BuildSys). ACM (2017)
16. Kotal, A., Piplai, A., Chukkapalli, S.S.L., Joshi, A.: PriveTAB: Secure and Privacy-Preserving sharing of Tabular Data. In: International Workshop on Security and Privacy Analytics (IWSPA). p. 35–45. ACM (2022)
17. Li, J., Cairns, B.J., Li, J., Zhu, T.: Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. NPJ Digital Medicine **6**(1), 98 (2023)
18. Liu, G., Tang, P., Hu, C., Jin, C., Guo, S., Stoyanovich, J., Teubner, J., Mamoulis, N., Pitoura, E., Mühlig, J.: Multi-dimensional data publishing with local differential privacy. In: EDBT. pp. 183–194 (2023)
19. Ma, C., Li, J., Ding, M., Liu, B., Wei, K., Weng, J., Poor, H.V.: RDP-GAN: A Rényi-Differential Privacy Based Generative Adversarial Network. IEEE Trans on Dependable and Secure Computing **20**(6), 4838–4852 (2023)

20. Majeed, A., Lee, S.: Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. IEEE Access **9**, 8512–8545 (2021)
21. Mironov, I.: Rényi Differential Privacy. In: IEEE Computer Security Foundations Symposium (CSF). pp. 263–275 (2017)
22. Mosquera, L., El Emam, K., Ding, L., , et al.: A method for generating synthetic longitudinal health data. BMC Medical Research Methodology **23**(1), 67 (2023)
23. Papernot, N., Abadi, M., Úlfar Erlingsson, Goodfellow, I., et al.: Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data (2017)
24. Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Úlfar Erlingsson: Scalable Private Learning with PATE. In: Int. Conf. on Learning Representations (ICLR) (2018)
25. Park, M., Foulds, J., Choudhary, K., Welling, M.: DP-EM: Differentially Private Expectation Maximization. In: International Conference on Artificial Intelligence and Statistics. vol. 54, pp. 896–904. PMLR (2017)
26. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: Symposium on Security and Privacy (IEEE S&P). pp. 3–18 (2017)
27. Song, L., Mittal, P.: Systematic Evaluation of Privacy Risks of Machine Learning Models. In: 30th USENIX Security Symposium. pp. 2615–2632 (2021)
28. Stadler, T., Oprisanu, B., Troncoso, C.: Synthetic Data – Anonymisation Groundhog Day. In: 31st USENIX Security Symposium. pp. 1451–1468. Boston, MA (2022)
29. Sun, C., van Soest, J., Dumontier, M.: Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. Journal of Biomedical Informatics **143**, 104404 (2023)
30. Tang, X., Mahloujifar, S., Song, L., Shejwalkar, V., Nasr, M., et al.: Mitigating Membership Inference Attacks by Self-Distillation Through a Novel Ensemble Architecture. In: 31st USENIX Security Symposium. pp. 1433–1450 (2022)
31. Torfi, A., Fox, E.A., Reddy, C.K.: Differentially private synthetic medical data generation using convolutional GANs. Information Sciences **586**, 485–500 (2022)
32. Wang, T., Yang, X., Ren, X., Yu, W., Yang, S.: Locally Private High-Dimensional Crowdsourced Data Release Based on Copula Functions. IEEE Trans on Services Computing **15**(2), 778–792 (2022)
33. Yoon, J., Drumright, L.N., van der Schaar, M.: Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). IEEE Journal of Biomedical and Health Informatics **24**(8), 2378–2388 (2020)
34. Yoon, J., Jordon, J., van der Schaar, M.: PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees. In: Int. Conf. on Learning Representations (ICLR) (2019)
35. Yoon, J., Mizrahi, M., Ghalaty, N.F., et al.: EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. NPJ Digital Medicine **6**(1), 141 (2023)
36. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., Gao, Y.: A survey on federated learning. Knowledge-Based Systems **216**, 106775 (2021)
37. Zhao, Z., Kunar, A., Birke, R., Van der Scheer, H., Chen, L.Y.: CTAB-GAN+: enhancing tabular data synthesis. Frontiers in Big Data **6** (2024)
38. Zhu, L., Liu, Z., Han, S.: Deep leakage from gradients. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)