JSL, 9(6), June 2024: 335- 345

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

JOURNAL OF SCIENTIFIC LETTERS
www.jslsci.com

SECURE SYNTHETIC DATA MODELING THROUGH DIFFERENTIAL PRIVACY AND DEEP GENERATIVE NETWORKS

Harer Savita Laxman

Research Scholar, Department of Computer Science Engineering, Vikrant University, Gwalior, M.P.

Dr. Shashank Swami

Research Supervisor, Department of Computer Science Engineering, Vikrant University, Gwalior, M.P.

ABSTRACT

The growing need for solutions that protect patients' privacy has shifted researchers' attention to the creation of synthetic medical data. Although real-world datasets are essential for training existing generative methods, access to these datasets is frequently limited. On the other hand, statistics on these datasets are easier to come by, but existing algorithms have a hard time producing tabular data using only statistical inputs. Secure synthetic tabular data modeling is demonstrated in this paper using a Differentially Private Conditional Tabular Generative Adversarial Network (DP-CTGAN). Incorporating calibrated noise and gradient clipping into the training procedure allows for differential privacy, which provides formal privacy assurances without drastically lowering data quality. The results show that DP-CTGAN is a trustworthy framework for privacy-sensitive machine learning applications and safe data publication because it strikes a good mix between the realistic representation of synthetic data, the practicality of machine learning, and the strict preservation of user privacy.

Keywords: Differential Privacy, DP-CTGAN, Synthetic Data.

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

I. INTRODUCTION

In order to intelligently supplement human-typical procedures and workflows, machine learning applied to relational tabular data is becoming more commonplace in modern businesses. The data science platform KAGGLE has surveyed data types and found that tabular data is the most prevalent in business and the second most common in academics. Meanwhile, scientists are praising synthetic data for its potential to solve many common data science problems, such as removing bureaucratic roadblocks to data access, eliminating important bottlenecks, and creating a "safe data space" for investigation. For example, a company could provide its employees with synthetic data to shield them from access to personal information about their friends or celebrities, or it could give synthetic data to outside consultants to mitigate risk in the case of an unintentional breach. Synthetic datasets can be tailored to meet specific requirements, such as for testing new tools or developing educational tutorials.

For datasets P(D) that have been collected in the last ten years, synthetic data has been generated by first sampling from a joint multivariate probability distribution that has been modelled. More intricate distributions are needed for more complicated datasets. For instance, hidden Markov models may be used to describe a sequence of events, or copulas could be used to model a set of non-linearly linked variables. However, existing distribution functions significantly limit the types of representations that may be utilised to build generative models, which in turn limits the accuracy of the synthetic data produced by these models.

Simultaneously, a number of academics in the field of statistical sciences have started to produce synthetic data using methods based on randomisation. The majority of these initiatives sought to facilitate data sharing while also protecting the privacy of the individuals whose information was included in the data (often survey takers). Generative adversarial networks (GANs) and their various extensions are attractive generative models because they promise to generate and manipulate images and natural languages, and because they offer flexible and efficient data representation.

II. REVIEW OF LITERATURE

Adiputra, I Nyoman Mahayasa et al., (2024) The class imbalance is one of the many problems with customer attrition data. One of the most common problems is class overlap,

E-mail: ***editor.jslsci888@gmail.com***

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

which occurs when data instances are identical across several classes. When training data has a lot of overlapping classes, it becomes more harder to forecast which customers will leave. We suggested a tabular GAN-based hybrid method, CTGAN-ENN, to deal with class overlap and imbalanced data in datasets that contain churning consumers. We utilised five independent datasets on customer attrition that were obtained from an open platform. One way to fix class imbalance using tabular GAN-based oversampling is with CTGAN, although it has a problem with class overlap. By combining CTGAN with the ENN under-sampling technique, we were able to circumvent the class overlap. Feature-based class overlaps were effectively reduced across all datasets using CTGAN-ENN. We tested CTGAN-ENN's performance across all machine learning algorithms. The results of our experiments demonstrate that CTGAN-ENN does well on the KNN, GBM, XGB, and LGB machine learning tasks related to client turnover prediction. We discovered that CTGAN-ENN outperformed other ML algorithms that employ cost-sensitive learning, popular over-sampling and hybrid sampling methods, and other sampling methods and algorithm-level approaches. To connect CTGAN and CTGAN-ENN, we give a laborious method. Lower time consumption was observed with CTGAN-ENN compared to CTGAN. In this work, we created a unique approach that efficiently handles various forms of imbalanced datasets and can be utilised to real-world customer churn prediction data.

Liu, Tongyu et al., (2023) Due to the development of big data, there is an urgent need for data release that safeguards people' privacy. Traditional methods of addressing this issue have failed to produce an acceptable compromise between data privacy and practical use. When it comes to using GANs for tabular data synthesis, there are a lot of different approaches that people in the database and machine learning communities have proposed to deal with this problem. Since there hasn't been a comprehensive evaluation of GAN-based methods vs conventional ways, we don't know why and how GANs could beat traditional strategies when it comes to synthesising tabular data. It is also difficult for practitioners to understand which components are crucial when building a GAN model for tabular data synthesis. In order to address this gap, our extensive experimental study investigates the use of GAN to tabular data synthesis. In this paper, we provide a unified GAN-based framework and detail several design options for its various components, such as neural network designs and testing methods. We provide optimisation methods for handling real-world problems during GAN training. We conduct extensive experiments to explore the design space and compare our findings with those of more traditional ways of data creation. Our comprehensive testing

E-mail: ***editor.jslsci888@gmail.com***

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

leads us to the conclusion that GAN has great promise for tabular data synthesis, and we offer recommendations for making the best design decisions. Additionally, we point out a few areas where GAN falls short and propose some directions for future research. Researchers of the future will be able to access and utilise all datasets and source code.

Shafqat, Wafa & Byun, Yungcheol. (2022). The growth of information technology is directly responsible for the explosion of data kept online. Recommendation systems have developed into an effective tool for data filtering, which helps with information overload. It is challenging for machine learning algorithms to build these recommendation systems since most real-world datasets are not balanced. In an unbalanced dataset, the classifier will under- or over-prioritize the majority class, leading to less accurate predictions. When it comes to sampling under-represented populations, there are several tried-and-true strategies that have stood the test of time. On the other hand, synthetic tabular data that closely mimics the original data while retaining its probability distribution may be effectively produced using generative adversarial networks (GAN). We provide a hybrid GAN approach to fix the data imbalance problem and boost recommendation system performance. The use of conditional Wasserstein GAN with gradient penalty allowed for the generation of tables that contained both numerical and categorical information. With the use of improved auxiliary classifier loss, we double-checked that the model reliably gave minority-class results. Based on the PacGAN concept, we constructed the discriminator architecture to receive m-packed samples instead of a single input. Our proposed model successfully addressed the mode collapse problem by utilising the PacGAN architecture. Two methods were used to assess our model. First, based on how well the generated data meets quality standards; and second, comparing the relative effectiveness of several recommendation models that were fed either the generated data or the original data.

Bourou, Stavroula et al., (2021) The availability of public data and advancements in technology have both resulted in a surge in attacks on computer systems. Networks are mostly reliant on Intrusion Detection Systems (IDS) for defence against malevolent actors. Machine learning approaches furnish a diverse selection of cybersecurity assets. However, for these strategies to be properly taught, extensive data sets are necessary, which can be problematic to obtain or employ because of privacy considerations. The Generative Adversarial Network (GAN) is a significant Machine Learning approach with encouraging applications in generating tabular data. To start, we offer a concise outline of the three most

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

prevalent GAN frameworks employed in this research: VanillaGAN, WGAN, and WGAN-GP. Models such CTGAN, CopulaGAN, and TableGAN are utilised for generating synthetic IDS data, with a particular emphasis on tabular information. Notably, with regard to the limits and expectations of this procedure, the models are trained and tested with an NSL-KDD dataset. We wrap up by supporting and analysing the foremost GANs for the generation of tabular network data utilising distinct quantitative and qualitative techniques.

III. EXPERIMENTAL SETUP

In order to evaluate the performance and utility compromise of the CTGAN model relative to its differentially private counterpart, we will utilise the benchmarking suite that was initially used to assess CTGAN. Created by MIT, SDGym2 functions as a benchmark for assessing the performance of synthetic data generators. It includes metrics to assess the creation of both single and many tables for machine learning and publication intents. Metrics comprise utility measurements from statistical and machine learning domains, and privacy metrics.

We implement CTGAN, our proposed DP-CTGAN, and a baseline model on various metrics using multiple datasets.

Data sets

We chose four commonly utilised tabular datasets from the UCI Machine Learning Repository and Kaggle to assess the efficacy of CTGAN and the suggested DP-CTGAN model. These datasets encompass a variety of sectors such as healthcare, finance, consumer analytics, and telecommunications, rendering them appropriate for evaluating synthetic data production within the framework of differential privacy. The datasets encompass a combination of continuous, binary, and categorical features, including both classification and regression assignments. The overarching traits are encapsulated in Table 1.

Table 1: Benchmark Dataset Characteristics

Dataset	Continuous Columns	Binary Columns	Categorical Columns	Records	Task
Heart Disease	7	5	2	1,025	Classification

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

Bank Marketing	10	6	5	45,211	Classification
Telecom Customer Churn	8	7	4	7,043	Classification
California Housing	8	0	1	20,640	Regression

Models

We evaluate the effectiveness and privacy compromise of the proposed DP-CTGAN model in comparison to a conventional CTGAN, utilising an Identity Generator (which reproduces the original dataset) as a reference model. To enhance the evaluation of the privacy scheme's efficacy, we will assume that the univariate attributes in our dataset are publicly accessible. The GANs will be trained utilising identical hyperparameters, employing a batch size of 500, and each model will undergo training for 100 epochs. For DP-CTGAN, we will utilise varying noise multipliers (nm = 10-5, 0.01, 0.1, 1) to get distinct privacy levels, and we will present the average epsilon of the models across the different datasets in the results section. The aim of this assessment is to demonstrate that we start with same CTGAN performance while introducing little noise (10-5) and conclude with a significant privacy constraint ($\varepsilon \sim 1$) upon the introduction of substantial noise. In all models and assessments, δ is consistently established at 10-3, signifying a substantial predicted resilience against hostile assaults within our privacy assurance.

Utility Metrics

Utilising our dataset, we have a machine learning assignment aimed at assessing the effectiveness of synthetic data production methods through machine learning to evaluate the value of the Synthetic Dataset. Figure 1 depicts the assessment framework. The machine learning algorithms employed for this assessment include Decision Trees, AdaBoost, Logistic Regression, and Multi-Layer Perceptron.

E-mail: ***editor.jslsci888@gmail.com***

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

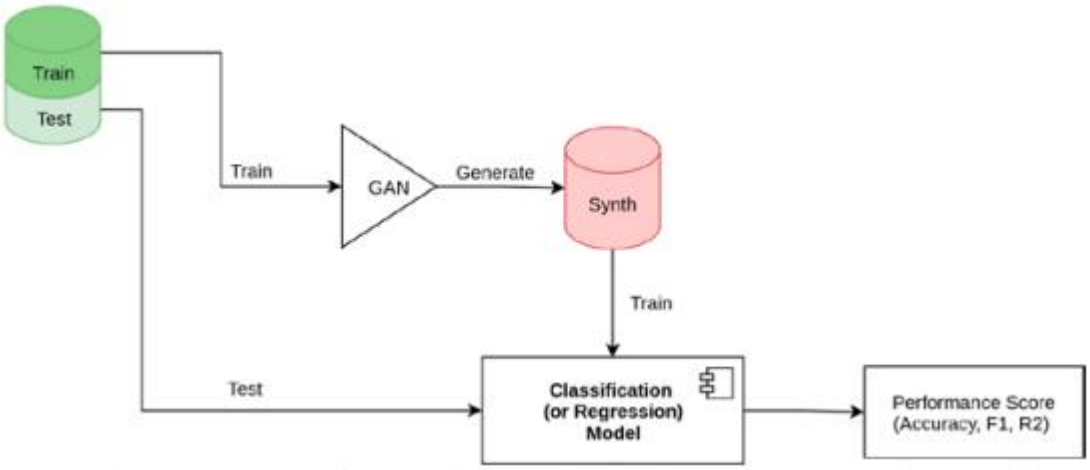

Figure 1: Utility of the synthetic dataset against the real dataset

Detection Metrics

We will employ a Detection methodology to assess the differentiability of the Real dataset from the synthetic one. The two selected measures construct a Machine Learning Classifier that distinguishes fake data from authentic data. The classifier undergoes assessment by Cross Validation, calculating one minus the mean ROC AUC value achieved. The selected machine learning models for this assessment are SVC Detection and Logistic Detection, as seen in Figure 2.

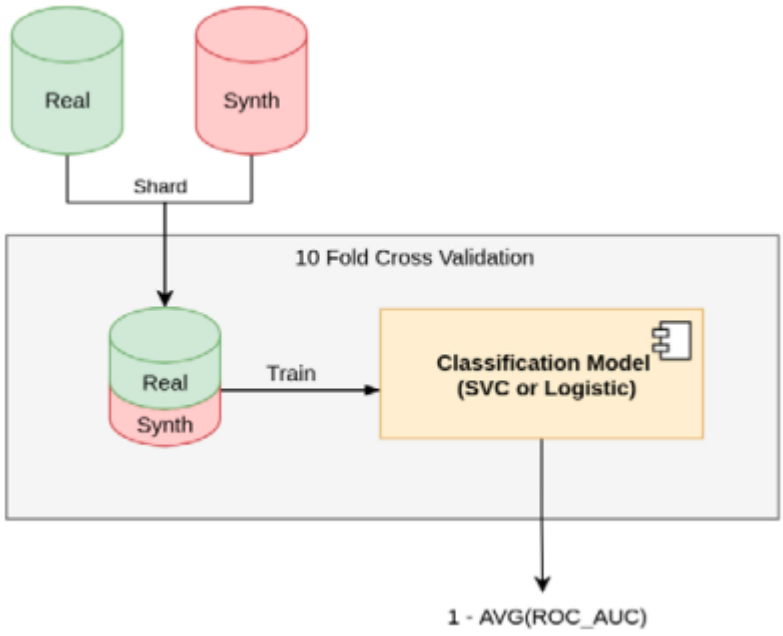

Figure 2: Distinguishability of synthetic data from the Real data

E-mail: editor.jslsci888@gmail.com

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

Privacy Metrics

For each dataset, we individually address the attributes by using SVM and Random Forest models to categorical characteristics, while utilising a Support Vector Regressor (SVR) and Linear Regression for numerical attributes. The assessment framework is illustrated in Figure 1.

A significant observation is that these metrics aim to evaluate the machine learning efficacy of synthetic data; nevertheless, in theory, any metric may be employed to contrast with alternative techniques for producing private tabular datasets.

IV. RESULTS AND DISCUSSION

We assessed our models utilising the aforementioned approach. The benchmark outcomes are encapsulated in Table 2. We categorised the outcomes in the table according to the objectives that the measurements seek to fulfil. We further categorise into 'Classification' and 'Regression' activities for machine learning use. We further differentiated between 'Numerical' and 'Categorical' measures in the context of privacy.

Table 2: Comparative Performance of CTGAN and DP-CTGAN Models under Different Privacy Levels

Method	Noise Multiplier	Utility (Classification)	Utility (Regression)	Detection	Privacy (Numerical)	Privacy (Categorical)	Epsilon (ε)
Identity	N/A	0.89	0.91	1.00	0.18	0.54	∞
CTGAN	N/A	0.84	0.82	0.76	0.35	0.63	∞
DP-CTGAN	0.00001	0.83	0.80	0.72	0.47	0.69	>100000
DP-	0.001	0.81	0.77	0.64	0.55	0.76	7524

E-mail: editor.jslsci888@gmail.com

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

CTGAN							
DP-CTGAN	0.10	0.78	0.69	0.39	0.67	0.88	81
DP-CTGAN	1.00	0.71	0.54	0.19	0.83	0.95	0.97

The benchmark datasets that were used to assess the proposed DP-CTGAN model included data from healthcare, banking, customer attrition, and housing. In order to enable reliable machine learning models, the experimental findings show that lower noise multipliers sustain better data usefulness while making synthetic datasets look very similar to the real data. Reduced distinguishability and higher resistance to attribute inference assaults are indicators of improved privacy protection as the noise multiplier grows. Despite a steady decline in classification accuracy and regression performance as privacy settings rise, the produced datasets nevertheless have sufficient analytical value for the majority of real-world uses. Results show that DP-CTGAN achieves a good trade-off between data usefulness and privacy by meeting differential privacy requirements while still providing synthetic tabular data that looks realistic.

V. CONCLUSION

Using differential privacy when training a model makes it more resistant to membership and attribute inference attacks and less likely that information would leak. Testing the suggested framework on industry-standard datasets in healthcare, finance, telecoms, and housing proved that it strikes a good compromise between data value and privacy. Under moderate privacy settings in particular, the produced synthetic datasets demonstrated dependable machine learning performance while preserving crucial statistical features. Overall, the data's analytical value was still enough for real-world applications, even if the predicted accuracy gradually declined as the privacy budget was increased. It is clear from the results that DP-CTGAN is a reliable, scalable, and powerful tool for research, machine learning, and privacy-sensitive data sharing.

E-mail: ***editor.jslsci888@gmail.com***

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

REFERENCES

Adiputra, I. N. M., Mahayasa, I. N., & Wanchai, P. (2024). CTGAN-ENN: A tabular GAN-based hybrid sampling method for imbalanced and overlapped data in customer churn prediction. *Journal of Big Data, 11*(1), 1–25.

Bourou, S., El Saer, A., Velivasaki, T.-H., Voulkidis, A., & Zahariadis, T. (2021). A review of tabular data synthesis using GANs on an IDS dataset. *Information, 12*(9), Article 375. https://doi.org/10.3390/info12090375

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM, 63*(3), 139–144. https://doi.org/10.1145/3422622

Haruna, Y., Qin, S., & Kiki, M. J. M. (2023). An improved approach to detection of rice leaf disease with GAN-based data augmentation pipeline. *Applied Sciences, 13*(3), 1346. https://doi.org/10.3390/app13031346

Iliyasu, A. S., & Deng, H. (2022). N-GAN: A novel anomaly-based network intrusion detection with generative adversarial networks. *International Journal of Information Technology, 14*(7), 3365–3375. https://doi.org/10.1007/s41870-022-00952-8

Liu, T., Fan, J., Li, G., Tang, N., & Du, X. (2023). Tabular data synthesis with generative adversarial networks: Design space and optimizations. *The VLDB Journal, 33*(2), 1–26. https://doi.org/10.1007/s00778-023-00807-y

Lu, H., Du, M., Qian, K., He, X., & Wang, K. (2021). GAN-based data augmentation strategy for sensor anomaly detection in industrial robots. *IEEE Sensors Journal, 22*(18), 17464–17474. https://doi.org/10.1109/JSEN.2021.3106510

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment, 11*(5), 1071–1083. https://doi.org/10.14778/3236187.3236211

Sampath, V., Maurtua, I., Aguilar Martín, J. J., & Gutierrez, A. (2021). A survey on generative adversarial networks for imbalance problems in computer vision tasks. *Journal of Big Data, 8*(1), 1–59. https://doi.org/10.1186/s40537-021-00504-8

Online ISSN: 2456-1495
CODEN (USA): JSLOBJ

Sattigeri, P., Hoffman, S. C., Chenthamarakshan, V., & Varshney, K. R. (2019). Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development, 63*(4–5), 3:1–3:13. https://doi.org/10.1147/JRD.2019.2941769

Sauber-Cole, R., &Khoshgoftaar, T. M. (2022). The use of generative adversarial networks to alleviate class imbalance in tabular data: A survey. *Journal of Big Data*, *9*(1), 98. https://doi.org/10.1186/s40537-022-00630-z

Scott, M., & Plested, J. (2019). GAN-SMOTE: A generative adversarial network approach to synthetic minority oversampling. *Australian Journal of Intelligent Information Processing Systems, 15*(2), 29–35.

Shafqat, W., & Byun, Y. (2022). A hybrid GAN-based approach to solve imbalanced data problem in recommendation systems. *IEEE Access*. PP(99),1-5.https://doi.org/10.1109/ACCESS.2022.3141776

Tanaka, F. H. K. D. S., & Aranha, C. (2019). Data augmentation using GANs. In *Proceedings of Machine Learning Research 30*(2), 1–16.

Yang, Z., Li, Y., & Zhou, G. (2023). TS-GAN: Time-series GAN for sensor-based health data augmentation. *ACM Transactions on Computing for Healthcare*, *4*(2), 1–21. https://doi.org/10.1145/3569473