Data Augmentation for Tabular Datasets Using Generative Adversarial Networks (GANs)

Ummaneni Dinesh Kumar[1], J Rajasekhar[2*], Pillala Lakshman[3] and Kolanti Manoj Kumar[4]
{2100050021@kluniversity.in[1], rajasekharemb@kluniversity.in[2*], 2100050066@kluniversity.in[3], 2100050037@kluniversity.in[4]}

Department of Electronics and Computer Science, Koneru Lakshmaiah Education Foundation, Green Fields, Vaddeswaram, Guntur Dist, Andhra Pradesh, India[1, 2, 3, 4]

Abstract. In the data hungry era of machine learning, the problem of datasets shortage, data imbalance, and privacy constraints impede the model effectiveness and ethical conformance. In this work, we propose a new approach to enhance data privacy with the help of the Generative Adversarial Networks (GANs). The method addresses concerns for protecting sensitive data and the need for augmentations to be effective. The proposed approach utilizes the synthetic data generation capability of GAN for increasing the diversity and representativeness of the training data whilst preserving privacy and avoiding the re-identification of actual records. The model uses a hybrid network with a conditional GAN (CGAN) to treat underrepresented classes together with a differential privacy (DP) approach to anonymize synthetic versions. Our framework is based on TensorFlow and focuses on adversarial learning for private realism without sacrificing model quality and privacy. Experiments on unbalanced benchmark data reveal that the generated data benefits later models (e.g., classification accuracy and F1-score), also resists membership inference attack, and can satisfy data protection regulation. This research fills a niche between data augmentation and ethical AI, for scalable solutions in domains such as healthcare and finance. Code, results and comparisons are provided to encourage reproducibility and following work.

Keywords: Generative Adversarial Networks, Data Augmentation, Data Imbalance, Privacy.

1 Introduction

Deep Learning models trained on sensitive data like medical history need to be trained in such a way that the privacy is maintained. One way is to train models on private data, and release only an indirect “student” readout model trained on public data that is not sensitive. The teacher model makes predictions on public data but introduces noise in a controlled manner to protect privacy. The student subsequently learns from these noisy predictions. Using this method, knowledge is transferred securely while maintaining privacy of individuals [1].

Privacy preserving mechanisms such as differential approach is used to obfuscate and anonymize the individual’s data to assure the reuse of the complete dataset and still protect the individual's privacy. Noise is intentionally generated in either the data or the predicted outputs, and sensitive inferences are hard to recover. According to the works in Green (2016) and Rajgarhia (2019), differential privacy provides a practical tradeoff between data utility and confidentiality2, which is of prime importance in sensitive domains like healthcare, finance and technology.

ICITSM-Part I 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2357951

In this work, we investigate how well Gans can be used for generating new data in several domains, including both image and table data. We analyze the influence of generalization samples in the performance of the model, understand processing difficulties when training Gans (mode collapse and instability) and some suggestions in the future as well. Our findings suggest GAN-based augmentation as a scalable and flexible solution that could lead to more reliable and accurate machine learning.

2 Literature Survey

Generative models, particularly generative adversarial networks (gans), are a wide range of techniques for creating realistic synthetic data. Gans introduced by Goodfellow et al. In 2014, two neuronal networks and identifiers consisted of a competitive debate process. This adversarial learning process will lead generators to provide more realistic samples, while also improving the ability of discriminators to distinguish between actual and synthetic data.

In this section, a selection of research studies that utilize the algorithms mentioned above are reviewed, and their findings are summarized:

TechFunnel (2020) [4] gives a broad summary of different anonymization approaches to augment how GANs contribute towards privacy and confidentiality in synthetic data generation. The creation of synthetic data is an important technology for modern scientists. M. Manohar [5] emphasizes value in solving the main problems faced by the actual data. One of the main advantages is the ability to process data, especially in the area where the actual data collection is expensive or powerful. Synthetic data also plays an important role in maintaining user confidentiality because it can imitate the actual data set without exposing confidential information. He also supports the production of more diverse and balanced data sets, increasing the performance and effectiveness of machine learning models. Overall, synthetic data generation enables the development of reliable and ethical AI systems.

Sharma and Bhatia (2017) [6] provides clear overview of various methods used to generate synthetic data to protect individual confidentiality. It explains all the old and new methods that help to retain meaningful patterns without identifying personal information. Some of these methods include artificial records that add noise to data or follow the same template as the actual template. Their review shows how to safely use synthetic data in areas such as medical and finance. In general, their work helps to understand how to balance the need for information that is useful for the confidentiality of the data.

The works [7] and [8] address the creation and exploitation of specific methods to generate synthetic data using the Generative Adversarial Networks (GANS). [7] Goodfellow et al. (2014), we consider the original GAN structure where the generator and the identifier play a game to generate data following the real distribution. The original idea is now realized in many generating models in other fields. 3 Results Justib XU and Martha Skoularidou Based on this study [8], XU and Skoularidou. (2019) introduces Conditional Tabular GANs (CTGAN), which is particularly derived for Table structure with numbers and categories. CTGAN also introduces conditional generation and normalisation approaches to alleviate imbalanced and complex data distributions. This study demonstrates the strength and flexibility of GAN in generating realistic synthetic data for a range of machine learning tasks. Yoon et al. (2019) [9] in which the idea of missing data imputation is incorporated together with GANs further validating our argument on

their versatility. Choi et al. (2017) [10], further applied GANs in the generation of multi-label discrete patient records, which is also able to capture evidence in the real data ofChoi et al. (2017) [10], expanded the use of GANs and demonstrated that GANs can be applicable to generate realistic multi-label discrete patient records, of GANs have been used in the healthcare domain to generate synthetic patient data preserving meaningful clinical information. This work shows that it is possible and feasible to use GANs to generate synthetic simulated data that can be used to simulate realistic medical data.

Investigate the use of Generative Adversarial Networks (GANs) to create synthetic table data [11] and [12]. These works show how GAN can be adapted and optimized to obtain realistic data suitable for machine learning models in structured data sets (e.g. existing GAN, Tgan). These approaches focus on preserving the statistical properties and feature relationships of the original data, making synthetic data highly valuable for analysis. This approach is especially useful when working with confidential or limited data sets because it provides safe and effective data exchange and analysis. Research also emphasizes problems such as the collapse of the regime and the instability of training that researchers continue to solve. In general, these contributions extend the boundaries of synthetic data production for actual application of machine learning. Xie et al. More advanced work by (2020) [13] explores differentially private GANs for tabular data, providing privacy in synthetic data generation while maintaining utility.

[14] uses an overload method based on GAN to focus on solving class imbalances in data sets. In many actual tasks of classification, certain classes cannot be protected, leading to biased models and prognosis performance. This study suggests the creation of synthetic samples for minority people to use GAN to balance the data set. While studying the distribution of minority classes, GANs generate realistic examples that help classifiers learn rare patterns. This increases the overall accuracy of the model and provides fair processing of all classes. This method is especially effective in areas such as fraud and medical diagnostic detection, where imbalance data is a common problem.

Both [15] and [16] serve as educational materials to simplify the basics of GAN and actual applications. Brownlee (2019) provides convenient introduction to beginners in the architecture and curriculum of the creation competition network, allowing people who are new to deep education can use the concept. Rashid 2021 discusses how to apply GANS in detail to expand this understanding and increase the data of the data tablet set. He explains how synthetic samples created by GANS can improve various data sets and improve the performance of the machine learning model. This papers together provides theoretical and practical information about the use of GAN for actual data tasks.

Both [17] and [18] emphasize the role of GAN in the production of synthetic data and confidential preservation. OpenAI (2020) emphasizes how to create a realistic synthetic data set so that HANS can imitate the statistical characteristics of the actual data so that the data can be replaced more safely without identifying confidential information. H2O.AI (2022) explores how GANs help to improve the productivity of machine learning models while maintaining confidentiality by increasing data as well as data generation. The two researchers emphasize the effect of the GAN -based method that improves the availability of data and supports reliable education of artificial intelligence models in confidential areas such as personal information protection and health care and finance.

3 Methodology

To evaluate the efficacy of Generative Adversarial Network (Gans) for data augmentation we implemented systemic technologies such as data harvesting, pre-processing, synthetic data generation, models training, and evaluation. 2 Methodology This section provides a brief overview of the main elements of the approach followed, namely: (i) the data preprocessing steps (preparation of spectral images of Sparrow sound), followed by (ii) the deep GAN architectures, and how they are employed in the context of data augmentation (spectrum augmentation).

3.1 Data Description and Preprocessing

We used the MIMIC-III database for length-of-stay prediction, which consists of 116,354 records and 27 attributes. This includes 26 categorical functions, including one continuous variable ("los"), "UDISSION_TYPE", "Insurance", "Paul" and various classification diseases, including a wide range of categories of demographic and diagnosis. The data set of MIMIC-III is used to predict mortality, with 58,976 lines and 19 columns, with 12 continuous and 7 categories. The dataset includes variables representing demographic, clinical, and administrative information, such as HADM_ID, Marital_status, Numdiagnoses, and Numlabevents. Missing values were handled using mean or median imputation depending on the attribute type. Standardization via Min-Max scaling was applied to numerical features to ensure stable GAN training. The dataset was split into training and testing sets in an 80:20 ratio to prevent data leakage and enable reliable model evaluation.

3.2 GAN-Based Synthetic Data Generation

We used various Gan architectures that matched the attributes of the data record. Some GAN options have been used to solve specific problems of data creation. Designed for table data, CTGAN effectively reflects complex functional dependence and includes mixed data types. The TGAN is used to promote the collapse of the mode and to increase the stability of learning, which causes more diverse and reliable synthetic data. Adam Optimizer trained the Gan model using Hyperparameter Tuning. This was refined using grid search to improve convergence. gans were trained for several eras until the loss of identification became stable, indicating that the identifiable and generator were in balance.

3.3 Model Training and Evaluation

The augmented datasets (comprising real and synthetic data) were utilized to train multiple machine learning models: Deep Learning Models, TGAN, CTGAN, and Multi-Layer Perceptron (MLP), XG Boost, Random Forest, KNN algorithm. We have compared the model output with the original, augmented data using evaluation metrics such as pairwise distances, distance matrix, and distance, Explained Variance Score, mean squared error, Mean Absolute Error, Root Mean Squared Error, Coefficient Determination.

3.3.1 Random Forest (RF)

Using a group of decision trees to increase resilience:

- Combines predictions from multiple trees using majority voting.
- Reduces overfitting compared to a single decision tree .

3.3.2 Gradient Boosting (e.g., XGBoost)

An iterative algorithm that optimizes weak learners (trees) by minimizing a loss function:

$$F\_(m+1)\ (x) = F\_m\ (x) + \gamma h\_m\ (x) \quad (1)$$

3.3.3 k-Nearest Neighbors (k-NN)

A distance-based algorithm:

- Predicts the class based on majority voting among k-nearest neighbors.
- Common distance metric**:** Euclidean Distance.

3.4 Model Evaluation

Performance metrics used to evaluate models:

- **Mean Absolute Error:**

$$(1/n) * \Sigma\ |yi - \hat{y}i| \quad (2)$$

- **Mean squared Error:**

$$(1/n) * \Sigma\ (yi - \hat{y}i)\text{^}2 \quad (3)$$

- **Root Mean Squared Error**:

$$\sqrt{([(1/n) * \Sigma\ (yi - \hat{y}i)\text{^}2\ ]\ )} \quad (4)$$

- **Explained Variance Score:** Explained variance measures prediction accuracy.
- **Coefficient Determination:** It shows how well predictions fit data
- **Lift$_i$:** within Distance$_i$ / external Distance

3.5 Workflow Summary

The process begins with the collection of raw data from relevant sources. Then follow preliminary processing. This includes data cleaning and preparation for modeling. Then, GAN (General Competitive Network) is trained as a pre -processed data for studying the template. After training, GAN is used in synthetic data to create data that mimics the actual data set. This synthetic data is then used in the educational model, which a machine learning model is created and specified. Fig 1 shows the work flow of this. Finally, the performance of this model was evaluated and analyzed to ensure accuracy and reliability.

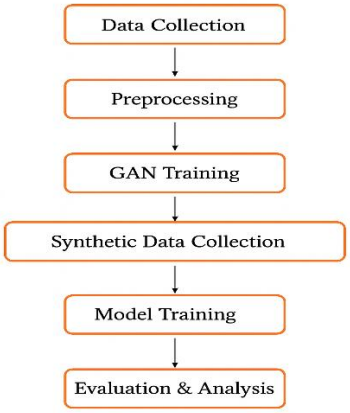

Fig.1. Work Flow.

3.6 Comparison of Results

From the evaluation of various models using both the original and synthetic datasets, we observed that the synthetic data performed differently from the original data in terms of error metrics. When teaching the source data from the XG boost model and random forest regression, we observed higher variance and dispersion, indicating a wider spread in the data. This typically correlates with more robust learning. However, the mean absolute error (MAE) for the synthetic data shows higher values, suggesting that the synthetic data has higher prediction errors compared to the original data. Nevertheless, when learning the synthetic data of TGA, their performance has worsened significantly, indicating that generalization is not good at the negative value of the described dispersion and R^2 evaluation. Meanwhile, CTGAN's synthetic data has been relatively better than the TGAN, especially in the return of XG Boost and any forest model. The KNN regression model has followed the similar trend of CTGAN's source data and the worst source data in TGAN. The neural network results showed minimal variation in R^2 values, but the error metrics demonstrated similar reductions when synthetic data was included. This indicates that while predictive variance remained consistent, synthetic augmentation contributed to lowering overall error. Table 1 shows the evaluation metrics.

Table 1. Evaluation Metrics.

Model Used	Data	Mean Absolute Error	Mean Squared Error	Root Mean Squared Error	Explained Variance Score	Coefficient of Determination
XG Boost Regression	Original	0.47	0.37	0.61	28.4%	28.4%
	TGAN Synthetic	0.73	0.79	0.89	-7.5%	-52.2%
	CTGAN Synthetic	0.54	0.47	0.68	9.7%	9.6%

Random Forest Regression	Original	0.49	0.40	0.63	23.2%	23.2%
	TGAN Synthetic	0.73	0.78	0.88	-4.3%	-48.7%
	CTGAN Synthetic	0.58	0.52	0.72	0.2%	0.2%
KNN Regression	Original	0.51	0.43	0.65	20.2%	18.6%
	TGAN Synthetic	0.77	0.87	0.93	-12.5%	-66.1%
	CTGAN Synthetic	0.57	0.53	0.72	-0.9%	-1.1%
Neural Network	Original	0.54	0.50	0.70	-	-
	TGAN Synthetic	0.77	0.88	0.94	-	-
	CTGAN Synthetic	0.68	0.78	0.88	-	-

In terms of Privacy at Risk (PAR), T-GAN reported a PAR of 1.02%, while CT-GAN reported 0.62%. In the approach to PAR (including distance indicators), T-GAN announced PAR 6.14%and CT-GAN-8.05%. We can see that PAR has now decreased to 0%. This is because the percentage of the TOP_REMOVE is set 100%by 100%, which deleted all synthetic data related to the initial record at risk. But the lack of confidentiality is not always meaningful to us. Table 2 represents the privacy at risk evaluation.

Table 2. Privacy at risk Evaluation.

S.No	**Privacy At Risk(par)**	**Par (with column matches)**	**Par (with distance matches)**	**Par (top remove percent is 100%)**
1	CT-Gan	0.62%	8.05%	0%
2	T-Gan	1.02%	6.14%	0%

4 Results

The model evaluation of the original and synthetic data sets (TGAN and CTGA) shows a noticeable difference in productivity and confidential risk. The models trained with initial data have consistently achieved the best results across all metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and explained variance. Among the synthetic data models, the data generated by CTGAN provides better results than the TGA, and the error is improved and the decrease in the râ² point is reduced, especially with the increase of XG and the regression of any forest. However, in terms of confidentiality, Privacy at Risk (PAR) analysis shows that TGAN preserved slightly more attribute-level privacy (1.02%) compared to CTGAN (0.62%).

Compared to the distance, CTGAN shows higher risk (8.05%) than TGAN (6.14%). Nevertheless, both in both cases, the ratio of the best records that must be deleted to protect the confidentiality remains at a level of 0%, indicating that the confidential data is not critically applied. In general, CTGAN maintains the risk of confidentiality within the allowable limits while creating a more realistic synthetic data that provides the best compromise between the utility and the confidentiality, providing a relatively better model productivity. On the contrary, TGAN sacrifices both performance and confidentiality, making CTGA a more favorable option among the two synthetic data generators.

5 Conclusion

The results of this project confirm that data augmentation using GANs provides an effective solution for improving machine learning models, particularly when data is limited or unavailable. GAN-based augmentation significantly improves model performance on tabular datasets by generating diverse synthetic samples. Despite challenges such as mode collapse and training instability, advanced architectures like TGAN and CTGAN have shown promising results in producing reliable synthetic data. Overall, GANs are demonstrated to be a valuable tool for addressing data scarcity and enhancing the robustness of machine learning models.

References

[1] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, ”Semi-supervised knowledge transfer for deep learning from private training data,” arXiv preprint arXiv:1802.06739, Feb. 2018.

[2] M. Green, ”What is differential privacy?” Cryptography Engineering Blog, Jun. 2016.

[3] A. Rajgarhia, ”A brief introduction to differential privacy,” Georgian Impact Blog, May 2019.

[4] TechFunnel, ”3 types of data anonymization techniques and tools to consider,” TechFunnel, Jul. 2020.

[5] M. Manohar, ”Synthetic data generation: A must-have skill for new data scientists,” Towards Data Science, Sep. 2020.

[6] D. Sharma and K. S. Bhatia, ”A review of synthetic data generation methods for privacy-preserving data publishing,” International Journal of Scientific Technology Research, vol. 6, no. 3, pp. 57-62, Mar. 2017.

[7] Goodfellow, I., et al. (2014). Generative Adversarial Nets. Advances in Neural Information Processing Systems (NeurIPS).

[8] Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K. (2019). Modeling Tabular Data using Conditional GAN. Advances in Neural Information Processing Systems (NeurIPS).

[9] Yoon, J., Jordon, J., van der Schaar, M. (2019). GAIN: Missing Data Imputation using Generative Adversarial Nets. International Conference on Machine Learning (ICML).

[10] Choi, E., et al. (2017). Generating Multi-Label Discrete Patient Records using Generative Adversarial Networks. Machine Learning for Health care Conference.

[11] Ashrapov, I. (2020). Tabular Data Generation with GANs. arXiv preprint arXiv:2010.00638.

[12] Lin, Z., et al. (2020). TabGAN: Modeling Tabular Data with Generative Adversarial Networks. IEEE BigData.

[13] Xie, L., et al. (2020). Differentially Private Generative Adversarial Network for Tabular Data. IEEE Transactions on Knowledge and Data Engineering.

[14] Kim, Y., et al. (2021). GAN-based Oversampling for Imbalanced Data Classification. Applied Sciences. 42.

[15] Brownlee, J. (2019). A Gentle Introduction to Generative Adversarial Networks (GANs). Machine Learning Mastery.

[16] Rashid, A. (2021). Data Augmentation in Tabular Data with GANs. Towards Data Science.

[17] OpenAI. (2020). Synthetic Data Generation with GANs.
[18] H2O.AI. (2022). Data Augmentation and Privacy with GANs. H2O Documentation9)