Synthetic Data Generation for Emergency Medical Systems: A Systematic Comparison of Tabular GAN Extensions

Md Faisal Kabir[a], Md Majharul Islam Nayem[b] and Sven Tomforde[c]
Institute of Computer Science, Department of Intelligent Systems, University of Kiel,
Christian-Albrechts-Platz 4, Kiel, Germany

Keywords: GAN, Medical Emergency, Data Generation, Synthetic Data, Deep Learning, Wasserstein LSTM-GAN, Data Privacy.

Abstract: The generation of synthetic medical data has gained significant attention due to privacy concerns and the limited availability of real medical datasets. Various methods and techniques have been employed across domains to address these challenges, especially for tabular data. This study presents a comparative analysis of multiple generative models and privacy concerns. In addition, we propose the WLSTM-GAN model, which is evaluated with and without privacy constraints specifically for three medical tabular datasets. Our model is designed to handle both categorical and continuous features independently, incorporating a single generator with two specialized LSTM networks, as well as two distinct discriminators tailored for continuous and categorical data. We demonstrate that LSTM-based architectures can be effectively adapted for tabular data generation, with our WLSTM-GAN outperforming several existing models in fidelity and privacy preservation.

1 INTRODUCTION

In recent years, various medical fields have undergone significant advancements; however, the emergency medical sector, particularly in prehospital care, still requires further updates and development (Piliuk and Tomforde, 2023). Prehospital environments frequently present numerous challenges and problems that must be addressed within a limited timeframe. Various emergency medical cases arise, requiring prompt attention from healthcare providers, including nurses and medical personnel (Kabir et al., 2024).

A particular instance of this is the telenotary. An AI system can support this by suggesting diagnoses and measures (Kabir and Tomforde, 2024). Training such an AI requires sophisticated medical data. Consider a scenario of emergency medical data with thirty-one unique diseases selected for analysis. Each case is associated with specific vital measurements and symptoms. Generating synthetic data that closely resembles real-world data could significantly enhance the accuracy of disease prediction or medication classification models.

For the generation of synthetic data, this paper compares several approaches: Generative Adversarial Networks (GAN) (Goodfellow et al., 2014), CTGAN (Xu et al., 2019b), DP-CTGAN (Fang et al., 2022), PETE-GAN (Jordon et al., 2018), PATE-CTGAN (Rosenblatt et al., 2020), TabFairGAN (Rajabi and Garibay, 2021), CopulaGAN (Rustad, 2022), CTabGAN+ (Zhao et al., 2022), our proposed Wasserstein LSTM-GAN (WLSTM-GAN) and including a variant with with differential privacy (DP) WLSTM-GAN. We further evaluate our approach using three medical datasets concerning specific healthcare applications, alongside the widely used adult dataset. Our primary objective is to generate synthetic data for medical applications.

The structure of this paper is organised as follows. Section 2 provides a review of the latest methods and data generation techniques. Section 3 describes the models employed in the study, with a detailed explanation of the proposed model, including data preprocessing steps, model architecture, and training procedures. Section 4 outlines the experimental framework, including the descriptions of the datasets and the evaluation metrics. Section 5 presents a comparative analysis of the models, along with a brief discussion. Finally, Section 6 concludes the study with key observations and final remarks.

[a] https://orcid.org/0009-0000-2354-2193
[b] https://orcid.org/0009-0006-2941-8058
[c] https://orcid.org/0000-0002-5825-8915

Kabir, M. F., Nayem, M. M. I. and Tomforde, S.
Synthetic Data Generation for Emergency Medical Systems: A Systematic Comparison of Tabular GAN Extensions.
DOI: 10.5220/0013307200003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 1199-1206
ISBN: 978-989-758-737-5; ISSN: 2184-433X

2 BACKGROUND: SYNTHETIC DATA GENERATION

Over the past several years, synthetic data generation has gained prominence over other scientific methodologies, including scientific and commercial domains (Surendra and Mohan, 2017). This rise in popularity is attributed to its advantages, including availability, scalability, improved data quality, and diversity. Over the past decade, deep learning models have become increasingly prominent for generating synthetic data. Deep learning is a specialised machine learning subarea that focuses on developing and performing complex learning tasks. Notable models include autoencoder (AE), GAN, autoregressive (AR), and diffusion models (DM).

AEs are an unsupervised learning technique well-suited for capturing complex, non-linear relationships within data. In non-linear domains, a prominent approach involves leveraging neural network architectures (Baldi, 2011). An autoencoder consists of an encoder and decoder connected by a bottleneck. The encoder compresses the input to transform it into a low-dimensional latent space, reconstructing the low-dimensional data representation back to the original input shape (Bank et al., 2021). The AE framework has several generative extensions, including variational (An and Cho, 2015), adversarial (Makhzani et al., 2016), Bayesian (Yong and Brintrup, 2022), and diffusion AEs. The aforementioned extended models are capable of generating new samples. Mean and standard deviation parameters characterize the latent space of the variational autoencoder, and the decoder learns the statistical distribution of the input data (Berahmand et al., 2024). However, VAE increases complexity and training time, with a notable challenge in producing high-quality samples. GAN offers a new approach to enhancing synthetic data generation to address this challenge. GAN is an unsupervised learning approach where two neural networks, generator and discriminator, interact. The generator produces synthetic data by sampling from random noise, and the discriminator assesses whether the data is authentic or generated data (Goodfellow et al., 2014). On the other hand, it increases the network complexity and creates a high chance of mode collapse. To solve the mode collapse problem, unrolled GAN (Metz et al., 2017) and Wasserstein GAN (Arjovsky et al., 2017) have been proposed. GAN are capable of handling diverse data modalities, including image (Goodfellow et al., 2014), text (Yu et al., 2017), audio (Donahue et al., 2019), and tabular (Xu et al., 2019a) data.

Synthetic data is utilized across a diverse range of sectors, including automotive and robotics, banking and finance, agriculture, eCommerce (Nadamoto et al., 2023), healthcare (Pezoulas et al., 2024), computer vision, audio processing, education (Vie et al., 2022), risk management, and manufacturing (Werda et al., 2024), among others (Berahmand et al., 2024).

3 APPROACH TO SYNTHETIC DATA GENERATION USING GANS

3.1 Data Generation Models

GAN. A GAN is structured by generator and discriminator models interacting through adversarial training. However, random noise serves as input to the generator and produces synthetic data, while the discriminator is trained to distinguish between real and synthetic data generated by the generator, outputting a probability score for each. The generator parameters are updated based on the discriminator's feedback. As a result, the generator produces more realistic synthetic data, while the discriminator enhances distinguishing between actual and synthetic data during training. Figure 1 depicts the architecture of the GAN model (Goodfellow et al., 2014).

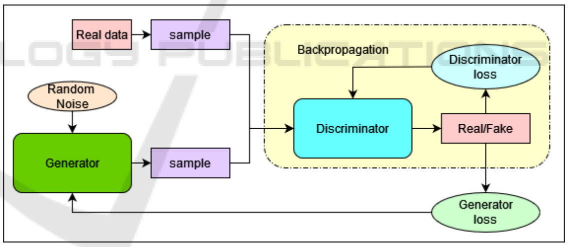

Figure 1: Generative adversarial network architecture.

CTGAN. Conditional tabular generative adversarial network (CTGAN) represents an extension of the Wasserstein GAN approach, particularly emphasising tabular-format synthetic data generation. CT-GAN proposed mode-specific normalization, particularly for continuous data types, addressing multimodal and non-gaussian distribution through mim-max normalization of the data range of [-1, 1]. Multivariate Gaussian distributions were employed before input into the GAN feed due to the diverse distributions among numerical features. The generator represents the conditional distribution of rows based on specific column values, hence termed a Conditional Generator, and the model a Conditional GAN (Xu et al., 2019b).

DP-CTGAN. Differentially private CTGAN (DP-CTGAN) represents an updated version of CTGAN.

Managing privacy loss from the generator or discriminator becomes increasingly complex, leading to a measurable degree of privacy leakage. In response to these issues, noise is applied solely to the discriminator and its parameter count is reduced to simplify the model. This leads to improved convergence and facilitates the estimation of privacy loss while the generator produces high-quality synthetic data with privacy protection (Fang et al., 2022).

PETE-GAN and PATE-CTGAN. PETE-GAN, an abbreviation for private aggregation of teacher ensembles, leverages differential privacy to produce synthetic data. The proposed network architecture features a singular generator alongside two separate discriminator blocks. The generator accepts the random noise as input and interfaces exclusively with the initial discriminator block referred to as the teacher discriminator. This block incorporates multiple discriminator units that classify the original and the generated data. The subsequent aggregation of the classifiers' votes determines the teachers' output. The resulting outputs from the student discriminator are essential for safeguarding the privacy of the original data (Jordon et al., 2018). PATE-CTGAN is an ensemble model approach with private aggregation of teacher ensembles and conditional tabular gan, named Quail-ified Architecture to Improve Learning (QUAIL). A teacher ensemble of GANs is trained on non-overlapping subsets of the real data, where each "teacher" model captures distinctive patterns from its respective data subset to generate synthetic samples (Rosenblatt et al., 2020).

TabFairGAN. TabFairGAN offers a new strategy to create synthetic tabular data using WGAN with fairness concerns by minimizing bias associated with sensitive attributes. This ensures that sensitive variables are processed fairly and equitably. Sensitive attributes, often treated as categorical features, can include binary or ordinal characteristics. The training of this network consists of two sequential phases: initially, synthetic data is generated, while the second phase introduces a fairness module, which allows the generator to refine its bias correction or learn the bias (Rajabi and Garibay, 2021).

CTabGAN+. CTabGAN+ is the updated version of CTabGAN that integrates several advancements, including a refined model architecture, enhanced support for different data types, and improved mechanisms for addressing missing values. CTabGAN+ demonstrates enhanced proficiency in addressing skewed distributions while featuring a more complex architectural design compared to its predecessor. CTabGAN+ employs a semi-supervised framework with a Wasserstein loss, which promotes more stable training and mitigates mode collapse. CTabGAN+ integrates differential privacy through stochastic gradient descent to safeguard data privacy (Zhao et al., 2022).

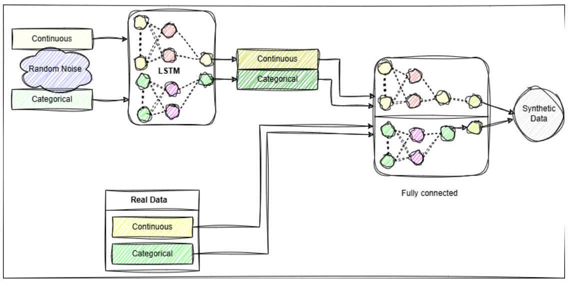

Figure 2: Wasserstein LSTM-GAN model architecture.

3.2 WLSTM-GAN

This study introduces two unique WLSTM-GAN frameworks employing Wasserstein loss, resulting in stable training and prevention of mode collapse. Additionally, models manage categorical and continuous features in distinct ways. The WLSTM-GAN(DP) s designed to address privacy issues using differential privacy. Both models incorporated long short-term memory (LSTM) networks solely for the generator, whereas fully connected layers are employed for two discriminators in both frameworks. In contrast, two separate discriminators process continuous and categorical features independently, sourcing inputs from real and generated data. Both networks adopt the Wasserstein loss and gradient penalty mechanism to promote training stability. Furthermore, we integrated differential privacy via stochastic gradient descent (DP-SGD) in the second model, ensuring stricter privacy protections during data synthesis.

3.3 Data Preparation

Let the dataset be represented by the variable $\mathbf{X}$, which represents $\mathbf{X}_i = (x_1, x_2, ..., x_N)$ consisting of $\mathbf{N}$ instances. x_i is a d-dimensional vector and each instance x_i contains d attributes and express as $x_i = (x_{i1}, x_{i2}, ..., x_{id})$. Here x_{ij} refers to the j-th attributes of the i-th observation. The attributes x_{ij} are categorized as continuous $\mathbf{X}_{con} \subset \mathbf{X}$ and categorical $\mathbf{X}_{cat} \subset \mathbf{X}$ features. A denoising LSTM autoencoder (Kabir and Tomforde, 2024) was utilized to input complete missing values in the continuous features, while the mode method was employed for the categorical features.

3.4 Model Architecture and Training

The WLSTM-GAN mechanism includes one generator G and two separate discriminators D_{con} and D_{cat}. In addition, generators and discriminators play a unique role. The G network is structured by LSTM layers and one fully connected (FC) output layer. Random noise is introduced as $Z_{\text{con}} \sim N(0,1)$ and $Z_{\text{cat}} \sim N(0,1)$ that signify continuous and categorical noise variables, respectively. The LSTM output is expressed as:

$$\text{LSTM}_{\text{out}} = \begin{cases} h_{\text{con}}, c_{\text{con}} = \text{LSTM}_{\text{con}}(Z_{\text{con}}) \\ h_{\text{cat}}, c_{\text{cat}} = \text{LSTM}_{\text{cat}}(Z_{\text{cat}}) \end{cases}$$

LSTM hidden layers are connected with FC layers, which represent synthetic continuous $\hat{X}_{\text{con}}$ and categorical $\hat{X}_{\text{cat}}$ data.

$$\text{FC}_{\text{out}} = \begin{cases} \hat{X}_{\text{con}} = w_{\text{con}} h_{\text{con}} + b_{\text{con}} \\ \hat{X}_{\text{cat}} = w_{\text{cat}} h_{\text{cat}} + b_{\text{cat}} \end{cases}$$

In the above expression, w, h, and b are denoted as weights, hidden state, and bias, respectively. So, the generator's input and output should be:

$$\text{G}_{(Z_{\text{con}}),(Z_{\text{cat}})} = \text{LSTM}_{Z_{\text{in}}} \to \text{LSTM}_{\text{out}} \to \text{FC}_{\text{in}} \to \text{FC}_{\text{out}}$$

$$\text{G}_{(Z_{\text{con}}),(Z_{\text{cat}})} = \hat{X}_{\text{con}}, \hat{X}_{\text{cat}}$$

Both discriminators are built on several fully connected layers with non-linear activations (Leaky ReLU). D_{con} accepts real continuous features $X_{\text{real(con)}}$ and synthetically generated features $\hat{X}_{\text{con}}$, which were generated by G and discriminates between them. On the other hand, the same process is performed for categorical features.

We designed the WLSTM-GAN, which requires a sophisticated training approach and stabilizes training using the Wasserstein loss with gradient penalty. The loss is defined as:

$$L_D = \mathbb{E}_{X_{\text{real}}}[D(X_{\text{real}})] - \mathbb{E}_{\hat{X}}[D(\hat{X})]$$

Here, L_D represent the loss function of the discriminator. $\mathbb{E}_{X_{\text{real}}}[D(X_{\text{real}})]$ and $\mathbb{E}_{\hat{X}}[D(\hat{X})]$ are the expected values of $D(X_{\text{real}})$ and $D(\hat{X})$ over real and generated data. Implementing the gradient penalty enforces the Lipschitz continuity condition on the discriminator by penalizing any variations from the desired gradient norm of 1. The interpolated sample X_{int}, the loss function for the gradient penalty L_{GP}, the weight controlling hyperparameter λ_{GP}, the expected value over the sampled points $\mathbb{E}_{X_{int}}$, the discrimination function $D(X_{int})$, applied to the interpolated data, and the deviation from the desired Lipschitz condition $\|\nabla_{X_{int}} D(X_{int})\|_2 - 1$ are defined as follows:

$$L_{GP} = \lambda_{GP} \mathbb{E}_{X_{int}} \left[(\|\nabla_{X_{int}} D(X_{int})\|_2 - 1)^2 \right]$$

4 EXPERIMENTAL DESIGN

4.1 Data Description

We use the following three open tabular data sets to compare the generative model sets from the medical domain: the emergency medical dataset, cardio dataset [1], diabetes dataset [2] and one common adult dataset [3]. All models, including ours, were trained on the adult dataset to ensure consistency in comparison. Our emergency medical dataset represents a proprietary collection of real-world medical data from Germany. Due to privacy concerns, it is not yet publicly available, though a partially anonymized version will be released soon.

Table 1: All datasets description.

Datasets	TS	TF	Con	Cat	Bi
Emergency	25k	30	8	22	0
Adult	32K	15	6	7	2
Cardio	70k	12	5	2	5
Diabetes	70k	22	4	3	15

Table 1 summarizes the characteristics of the datasets, including total samples (TS), total features (TF) and feature types (continuous, categorical, and binary) they contain.

4.2 Model Configuration

Optimizing hyperparameters plays a critical role in model tuning. This study incorporates three models: a generator with two LSTM networks, each associated with a fully connected layer. In addition, two discriminator models are employed to handle continuous and categorical data, respectively. To simplify model complexity, we fix the hyperparameters of the discriminators, excluding the learning rate.

Hyperparameter optimization was conducted using the Ray Tune Python library, with selected hyperparameters summarized in Table 2. These include learning rate (0.000001 to 0.01), batch size (16 to 512), and the generator model's LSTM hidden layers (0 to 4) and units (50 to 300). The LSTM output connects to a fully connected (FC) layer with hidden units ranging from 50 to 300. For the discriminator models, hidden layer configurations span 1 to 3 layers, with units ranging from 16 to 200.

[1] https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset (last accessed 2025-01-12)

[2] https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset (last accessed 2025-01-12)

[3] https://archive.ics.uci.edu/dataset/2/adult (last accessed 2025-01-12)

Table 2: Model hyperparameters and hyperparameter range.

Hyperparameters	value range
Learning rate	$0.000001 \leq n \leq 0.01$
Batch size	$16 \leq n \leq 512$
lstm hidden layers	$0 \leq n \leq 3$
lstm hidden units	$50 \leq n \leq 300$
FC units	$50 \leq n \leq 300$
Dis. hidden layers	$1 \leq n \leq 3$
Dis. hidden units	$16 \leq n \leq 200$

4.3 Evaluation Metrics

Kolmogorov-Smirnov (K-S) & Total Variation Distance (TVD) Test. The K-S (Jiang and Li, 2024) test is a nonparametric method widely employed in statistical analysis to compare cumulative distribution functions (CDFs) (Lautrup et al., 2024). The K-S test is employed for continuous data, while Total Variation Distance (TVD) (Lautrup et al., 2024) is applied to categorical features. TVD assesses distribution shape and location similarity using probability mass functions f_1 and f_2, independent of category order between populations. The TVD function is defined as follows:

$$\mathrm{TVD}(f_1, f_2) = \frac{1}{2}\sum_{x} |f_1(x) - f_2(x)|$$

Column Shape and Column Pair Trends. Column shape is a statistical methodology used to analyse the distribution of each column in generated datasets and compare it with the distribution observed in actual data. Column pair trends (sdm, 2023) investigate the correlation between original and generated continuous features using Pearson or Spearman coefficients. The contingency similarity between the real and synthetic categorical features is measured using the total variation distance and Cramer's V (Lautrup et al., 2024).

Data Validity. Data validity involves verifying that the generated data maintains uniqueness and that continuous features remain within the minimum and maximum ranges established by the original dataset. For categorical attributes, validity checks confirm that all generated categories correspond accurately to those in the original data (sdm, 2023).

NNDR & DCR. A quantitative measure evaluates the pointwise distance between synthetic and real samples by determining each synthetic sample's nearest-neighbor distance to real samples, comparing it against the distance to the subsequent nearest neighbor (Lautrup et al., 2024). A higher value indicates better performance, although it may imply a more significant potential privacy risk (Ooko et al., 2021). The DCR metric assesses the ratio of two median distances: the median of the distances from synthetic samples to their nearest real counterparts and the median of distances among real samples to their closest neighboring real samples. This ratio provides insight into how closely synthetic data mirrors real data in local structure, with implications for privacy depending on the degree of alignment between the two distributions (Zhao et al., 2021).

5 RESULTS AND ANALYSIS

This section describes our experimental setup for evaluating model performance on medical datasets, considering both data fidelity and data privacy constraints.

5.1 Comparative Analysis of Models

Data Fidelity. In tables 3 and 4 from the appendix, across the four datasets—Adult, Cardio, Emergency Medical, and Diabetes—TabFairGAN and WLSTM-GAN demonstrate exceptional performance, consistently achieving high scores in all three metrics. TabFairGAN exhibits particularly robust results in the Diabetes dataset, achieving near-perfect scores (CS = 0.983, CPT = 0.965, DV = 0.99), while WLSTM-GAN also behaves very well in all datasets. These results underscore their ability to accurately replicate statistical distributions, maintain feature correlations, and ensure data validity. In comparison, CTGAN and Copula GAN also perform well, particularly in CPT and DV metrics, where they achieve scores above 0.75 and close to 1.00, respectively, across most datasets. However, their CS scores are slightly lower than those of TabFairGAN and WLSTM-GAN, indicating potential limitations in fully replicating the column-wise statistical characteristics of the original data. PATE-GAN and PATE-ctGAN, while showing strong data validity (DV up to 1.00), struggle with column pair trends, with PATE-GAN having the lowest CPT scores between datasets (e.g., CPT = 0.028 for the Adult dataset). These models may not adequately capture inter-column dependencies, even when individual column distributions and data validity are preserved.

Traditional GAN models display notable weaknesses across all metrics, particularly in DV, where scores remain below 0.16. This underscores significant challenges in maintaining the uniqueness and consistency of generated data. The performance of DPctGAN and DP WLSTM-GAN, though stronger than traditional GANs, remains intermediate, with DV scores approaching 1.00 but lower CS and CPT scores compared to the top-performing models. These findings collectively underscore the importance of balanced performance across CS, CPT, and DV metrics in synthetic data generation. Models such as TabFairGAN and WLSTM-GAN, which excel in all three areas, are particularly well-suited for applications requiring high fidelity to the original data's statistical properties and interdependencies.

Data Privacy. The newRowSynthesis metric quantifies the number of rows in the synthetic dataset that are exact duplicates of the original data. A value of 1 indicates that no rows from the original dataset have been replicated in the synthetic data. The PATE-GAN, PATE-ctGAN, WLSTM-GAN, and DP WLSTM-GAN models consistently achieve a height value of 1 across all four datasets, demonstrating no duplication. In contrast, the TabFairGAN and CTabGAN+ models exhibit some degree of data duplication, as their scores are consistently near, but not equal to one. The DCR (Distance to Closest Record) and NNDR (Nearest Neighbor Distance Ratio) metrics compute specific measures for both the original and synthetic datasets. DCR evaluates the median distance to the closest record, while NNDR assesses the ratio of distances to the nearest neighbor. When the values for these metrics are similar between the original and synthetic datasets, it suggests a high likelihood of replicating the original data, thereby compromising data privacy. The degree of divergence between the artificial and original data is controlled by the model parameter delta. In this context, the PATE-GAN, PATE-ctGAN, DP WLSTM-GAN, and CTGAN models demonstrate greater privacy preservation, generating synthetic datasets that effectively protect the original data.

5.2 Discussion

Across multiple datasets (Adult, Emergency Medical, Cardio, and Diabetes), TabFairGAN, WLSTM-GAN, and CTabGAN+ consistently demonstrate strong data generation capabilities, achieving high scores in metrics assessing distributional fidelity, column shape and inter-variable relationships, which are essential for generating reliable synthetic data. Specifically, these models excel in column shape and trend replication, maintaining high data validity, thus closely approximating original datasets in univariate and multivariate dependencies. Our proposed model also dominates some of the available models. In privacy evaluations, however, models such as GAN and PATE-GAN often show elevated metrics in DCR and NNDR, suggesting higher re-identification risks. At the same time, TabFairGAN, CTabGAN+, and CopulaGAN typically exhibit lower DCR values, indicating more robust privacy protection. The results suggest that while TabFairGAN, WLSTM-GAN, and CTabGAN+ offer promising options for high-fidelity data generation, privacy safeguards, particularly in models such as GAN and PATE-GAN, require further refinement to ensure secure and privacy-compliant synthetic data applications, especially in sensitive domains. Incorporating noise into the WLSTM-GAN model to enhance privacy is problematic, as even minimal noise disrupts the original data's range, compromising data integrity. This alteration undermines the model's utility for privacy preservation within the WLSTM-GAN framework, as it may impair the model's ability to represent the original data distribution accurately.

6 CONCLUSIONS

Generating synthetic tabular data for medical applications is challenging due to the complex dependencies and specific value ranges required for accurate disease prediction and classification. Our proposed model, WLSTM-GAN, addresses these challenges by employing a single generator with two distinct networks to handle categorical and continuous variables alongside two separate discriminators tailored for categorical and continuous features. This study demonstrates that LSTM networks are not only effective for generating time-series data but also adaptable for tabular data generation. Furthermore, we conduct a comprehensive performance comparison of multiple state-of-the-art models, both with and without privacy-preserving mechanisms, using diverse datasets that include adult demographic data, emergency medical records, cardiovascular health data, and diabetes-related metrics. This evaluation enables a nuanced analysis of each model's efficacy and privacy-preserving capability across different types of medical and demographic information types, highlighting their adaptability and reliability within varied healthcare and demographic applications.

ACKNOWLEDGEMENTS

This work was funded by the German Ministry of Education and Research (Bundesministerium für Bildung und Forschung) as part of the project “KI-unterstützter Telenotarzt (KIT2)” under grant number 13N16402. We would like to thank them for their support.

REFERENCES

(2023). *Synthetic Data Metrics*. DataCebo, Inc. Version 0.12.0.

An, J. and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. arXiv:1701.07875 [cs, stat].

Baldi, P. (2011). Autoencoders, unsupervised learning and deep architectures. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop - Volume 27*, UTLW'11, pages 37–50. JMLR.org.

Bank, D., Koenigstein, N., and Giryes, R. (2021). Autoencoders.

Berahmand, K., Daneshfar, F., Salehi, E. S., and Li, Y. (2024). Autoencoders and their applications in machine learning: a survey. 57(2):28.

Donahue, C., McAuley, J., and Puckette, M. (2019). Adversarial audio synthesis.

Fang, M. L., Dhami, D. S., and Kersting, K. (2022). DP-CTGAN: Differentially private medical data generation using CTGANs. In Michalowski, M., Abidi, S. S. R., and Abidi, S., editors, *Artificial Intelligence in Medicine*, volume 13263, pages 178–188. Springer International Publishing. Series Title: Lecture Notes in Computer Science.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., and Xu, B. (2014). Generative adversarial networks.

Jiang, B. and Li, P. (2024). Nonparametric crosstalk evaluation method using the kolmogorov-smirnov test. In *2024 IEEE International Symposium on Electromagnetic Compatibility, Signal & Power Integrity (EMC+SIPI)*, pages 617–622. ISSN: 2158-1118.

Jordon, J., Yoon, J., and Schaar, M. v. d. (2018). PATE-GAN: Generating synthetic data with differential privacy guarantees.

Kabir, M. and Tomforde, S. (2024). A deep analysis for medical emergency missing value imputation. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence*, pages 1229–1236. SCITEPRESS - Science and Technology Publications.

Kabir, M. F., Aust, T., Cui, H., Botache, D., and Decke, J. (2024). *Organic Computing*. kassel university press.

Lautrup, A. D., Hyrup, T., Zimek, A., and Schneider-Kamp, P. (2024). SynthEval: A framework for detailed utility and privacy evaluation of tabular synthetic data.

Makhzani, A., Shlens, J., Jaitly, N., and Goodfellow, I. (2016). Adversarial autoencoders.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. (2017). Unrolled generative adversarial networks.

Nadamoto, A., Fukumoto, K., Takeuchi, R., and Terada, H. (2023). Automatic generation of product descriptions using deep learning methods. pages 134–148.

Ooko, S. O., Mukanyiligira, D., and Munyampundu, J. P. (2021). Synthetic exhaled breath data-based edge AI model for the prediction of chronic obstructive pulmonary disease. In *2021 International Conference on Computing and Communications Applications and Technologies (I3CAT)*, pages 1–6.

Pezoulas, V. C., Zaridis, D. I., and Mylona, E. (2024). Synthetic data generation methods in healthcare: A review on open-source tools and methods. 23:2892–2910.

Piliuk, K. and Tomforde, S. (2023). Artificial intelligence in emergency medicine. A systematic literature review. *Int. J. Medical Informatics*, 180:105274.

Rajabi, A. and Garibay, O. O. (2021). TabFairGAN: Fair tabular data generation with generative adversarial networks.

Rosenblatt, L., Liu, X., Pouyanfar, S., Leon, E. d., Desai, A., and Allen, J. (2020). Differentially private synthetic data: applied evaluations and enhancements.

Rustad, A. (2022). tabGAN: A framework for utilizing tabular GAN for data synthesizing and generation of counterfactual explanations. Master's thesis, NTNU.

Surendra, H. and Mohan, H. (2017). A review of synthetic data generation methods for privacy preserving data publishing.

Vie, J.-J., Rigaux, T., and Minn, S. (2022). Privacy-preserving synthetic educational data generation. In Hilliger, I., Muñoz-Merino, P. J., De Laet, T., Ortega-Arranz, A., and Farrell, T., editors, *Educating for a New Future: Making Sense of Technology-Enhanced Learning Adoption*, pages 393–406. Springer International Publishing.

Werda, M. S., Taibi, H., Kouiss, K., and Chebak, A. (2024). Generation of synthetic data for deep learning in manufacturing quality control systems. In *2024 IEEE 22nd Mediterranean Electrotechnical Conference (MELECON)*, pages 74–79. ISSN: 2158-8481.

Xu, L., Skoularidou, M., and Cuesta-Infante, A. (2019a). Modeling tabular data using conditional GAN.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019b). Modeling tabular data using conditional GAN.

Yong, B. X. and Brintrup, A. (2022). Bayesian autoencoders with uncertainty quantification: Towards trustworthy anomaly detection. 209:118196.

Yu, L., Zhang, W., Wang, J., and Yu, Y. (2017). SeqGAN: Sequence generative adversarial nets with policy gradient.

Zhao, Z., Kunar, A., Birke, R., and Chen, L. Y. (2022). CTAB-GAN+: Enhancing tabular data synthesis.

Zhao, Z., Kunar, A., Scheer, H. V. d., and Birke, R. (2021). CTAB-GAN: Effective table data synthesizing.

APPENDIX

The tables below display the results of data fidelity and privacy metrics for multiple dataset comparisons across various models, including our proposed approach. In DCR, R means real, and S means Synthetic.

Table 3: Adult and cardio dataset fidelity and privacy result regarding models.

Matrics →	adult fidelity			adult privacy			cardio fidelity			cardio privacy		
Model ↓	CS	CPT	DV	NRS	DCR	NNDR	CS	CPT	DV	NRS	DCR	NNDR
GAN	0.798	0.016	0.00	1.00	R:0.00007 S:171.9	R:0.042 S:0.504	0.000039	0.146	0.16	1.0	R:0.013 S:426	R:0.20 S:0.97
CTGAN	0.684	0.535	0.810	1.00	R:0.00006 S:0.00026	R:0.045 S:0.105	0.906	0.730	1.00	0.99	R:0.014 S:0.078	R:0.200 S:0.454
DP-CTGAN	0.416	0.289	1.00	1.00	R:0.00007 S:0.00052	R:0.040 S:0.031	0.752	0.633	0.99	1.00	R:0.014 S:2.298	R:0.208 S:0.484
CopulaGAN	0.735	0.528	0.800	1.00	R:0.00006 S:0.00016	R:0.037 S:0.071	0.924	0.770	1.00	0.99	R:0.012 S:0.018	R:0.214 S:0.244
TabFairGAN	0.935	0.874	0.99	0.99	R:0.00006 S:0.00014	R:0.039 S:0.108	0.966	0.843	1.00	0.96	R:0.012 S:0.025	R:0.222 S:–
CTabGAN+	0.882	0.857	0.99	0.99	R:0.00006 S:0.00016	R:0.038 S:0.062	0.975	0.793	1.00	0.96	R:0.012 S:0.025	R:0.200 S:–
PATE-GAN	0.929	0.028	0.182	1.00	R:0.00008 S:243	R:0.039 S:0.588	0.083	0.128	0.25	1.00	R:0.013 S:359	R:0.222 S:0.941
PATE-CTGAN	0.654	0.435	1.00	1.00	R:0.00007 S:0.13820	R:0.039 S:0.202	0.748	0.619	1.00	1.00	R:0.013 S:7.246	R:0.200 S:0.360
WLSTM-GAN	0.899	0.864	0.99	1.00	R:0.00006 S:0.00086	R:0.042 S:0.055	0.937	0.898	1.00	1.00	R:0.012 S:18.8	R:0.222 S:0.483
DP WLSTM-GAN	0.665	0.385	0.906	1.00	R:0.00007 S:0.13820	R:0.039 S:0.202	0.856	0.712	1.00	1.00	R:0.013 S:3.726	R:0.200 S:0.565

Table 4: Emergency medical and diabetes datasets fidelity and privacy result regarding models.

Matrics →	emer. medical fidelity			emer. medical privacy			diabetes fidelity			diabetes privacy		
Model ↓	CS	CPT	DV	NRS	DCR	NNDR	CS	CPT	DV	NRS	DCR	NNDR
GAN	0.020	0.061	0.133	1.00	R:0.033 S:105	R:0.300 S:1.799	0.014	0.003	0.045	1.00	R:0.051 S:13.60	R:0.500 S:0.893
CTGAN	0.863	0.759	1.00	0.92	R:0.034 S:0.078	R:0.587 S:0.627	0.845	0.779	0.95	1.00	R:0.055 S:0.111	R:0.500 S:0.502
DP-CTGAN	0.696	0.628	0.97	1.00	R:0.034 S:1.965	R:0.108 S:0.384	0.907	0.794	1.00	0.936	R:0.058 S:0.058	R:0.205 S:–
CopulaGAN	0.863	0.752	1.00	1.00	R:0.032 S:0.018	R:0.414 S:0.293	0.821	0.770	0.954	1.00	R:0.050 S:0.117	R:0.317 S:0.528
TabFairGAN	0.895	0.881	1.00	0.962	R:0.0376 S:0.025	R:0.292 S:0.493	0.983	0.965	0.99	0.97	R:0.058 S:0.058	R:0.425 S:–
CTabGAN+	0.765	0.586	1.00	0.919	R:0.032 S:0.025	R:0.341 S:0.352	0.955	0.914	1.00	0.977	R:0.058 S:0.117	R:0.500 S:–
PATE-GAN	0.071	0.072	0.09	1.00	R:0.03 S:73.6	R:0.134 S:0.852	0.070	0.012	0.130	1.00	R:0.058 S:16.6	R:0.500 S:0.952
PATE-CTGAN	0.836	0.674	0.98	1.00	R:0.038 S:0.246	R:0.199 S:0.394	0.806	0.714	1.00	1.00	R:0.058 S:0.529	R:0.500 S:0.479
WLSTM-GAN	0.890	0.677	0.96	1.00	R:0.032 S:0.829	R:0.478 S:0.492	0.985	0.9419	0.973	1.00	R:0.055 S:0.239	R:0.457 S:0.593
DP-WLSTM-GAN	0.796	0.513	0.953	1.00	R:0.033 S:3.726	R:0.203 S:0.358	0.765	0.736	0.93	1.00	R:0.058 S:0.129	R:0.500 S:0.438