2024 IEEE 4th International Conference on Digital Twins and Parallel Intelligence (DTPI) | 979-8-3503-4925-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/DTPI61353.2024.10778801

Design and Application of the Digital Human Identity Attribute Data Synthesis System

Yapeng Ji*†, Yangyang Li†, Yunji Liang*, Bin Guo*, Yangzhao Yang†
*School of Computer Science, Northwestern Polytechnical University, Xi'an, China
nwpu_jyp@mail.nwpu.edu.cn, liangyunji@nwpu.edu.cn, guob@nwpu.edu.cn
†Academy of Cyber, Beijing, China
liyangyang@ict.ac.cn, forester@mail.ustc.edu.cn

***Abstract*—In the era of big data, increasing demand for identity data sharing in fields like healthcare and tourism, along with stricter cybersecurity regulations, has made generating realistic digital human identity attribute data and ensuring its efficient use a critical challenge. However, existing methods struggle with complex data types and distributions, often neglecting the social and logical consistency of the synthetic data. To address this, we propose a Conditional Generative Adversarial Network (CGAN)-based system for synthesizing digital human identity attribute data. The system features three modules: data preprocessing, data synthesis, and data verification. The preprocessing module uses a composite encoder to handle discrete, continuous, and mixed data types, with a Variational Gaussian Mixture Model (VGM)-based normalization for continuous variables. The synthesis module introduces conditional vector-based sampling to resolve data imbalance while preserving data distribution characteristics, enhancing the realism of the generated data. The verification module uses association rule mining for secondary validation, ensuring logical consistency. This modular design improves scalability, paving the way for multimodal data synthesis. Experimental results show that the system excels in statistical similarity and machine learning utility, with the verification module boosting rule retention by 15%, significantly improving data quality and practical value. This study provides an effective solution with important theoretical and practical implications for digital human identity attribute data synthesis.**

***Index Terms*—CGAN,Data Synthesis,Digital Human Identity Attributes,Association Rule Mining,Tabular Data**

I. Introduction

With the advent of the big data era, the level of digitalization in society is continuously deepening, and personal identity information is increasingly transitioning to digital forms. Digital human identity attribute data, as a crucial resource, finds extensive applications in social media, healthcare, finance, and other fields. However, the widespread use of this digital information has raised societal concerns regarding privacy and security. Amid the conflict between the demands of various industries for application and the need for privacy protection of digital information, data synthesis technology has emerged as an effective solution to address this issue.

Data synthesis technology has achieved widespread application in many sensitive fields. For instance, in the healthcare sector, synthetic data can be used to generate virtual patient records, enabling the training of machine learning models to predict disease risk and diagnose illnesses without compromising privacy [1]. In the financial industry, synthetic data can generate virtual transaction records, assisting banks and financial institutions in training models to detect anomalies and potential fraudulent activities [2]. In the marketing field, synthetic data can create virtual consumer data, helping companies optimize their market strategies [3]. Specifically, the pseudo-data obtained through synthesis technology, which retains the characteristics of real data, is widely utilized in these application scenarios to support data-sharing needs under privacy protection constraints. However, existing methods still face numerous challenges in handling complex data types and imbalanced data distributions [4]. Moreover, they do not finely consider whether some synthetic data might contradict objective facts, such as generating an attribute pair like *"occupation=airline stewardess, gender=male."* This study refers to such situations as a lack of social logical consistency.

To address the aforementioned challenges, this study proposes a digital human identity attribute data synthesis system based on CGAN, with a specific focus on designing and optimizing the synthesis of tabular data—a common data format. The system is composed of the following key modules:

- **Data Preprocessing Module**: Initially, the module addresses issues such as missing and erroneous values in the raw data through cleaning processes. The cleaned composite data types are then categorized and encoded accordingly. For continuous data variables, a Mode-Specific Normalization based on the VGM or general transformations are employed to efficiently capture their distribution characteristics. Discrete data are processed using one-hot encoding. For mixed data types, they are treated as a combination of the aforementioned types. Finally, the encoded vectors are concatenated to achieve a unified encoding of various data types.
- **Data Generation Module**: Building on the unified encoding of data types from the previous module, the core network, GAN, is capable of fully learning the distribution of mixed vectors and their inter-column relationships. The introduction of fair sampling training based on conditional vectors effectively addresses the issue of class collapse in synthetic data caused by imbalanced

Corresponding author: Yangyang Li (liyangyang@ict.ac.cn) real data distribution, ensuring adequate learning of rare samples and ultimately generating high-quality synthetic data.

- **Data Verification Module**: To achieve more fine-grained validation of synthetic data quality and further improve its quality and usability, this module employs association rule mining to extract high-confidence rules from the real dataset. Based on these rules, secondary verification and correction are applied to the synthetic data, thereby enhancing its logical consistency.

The identity attribute data synthesis system proposed in this study not only improves the statistical similarity and machine learning utility of the generated data to a certain extent but also effectively maintains the social logical consistency of the data. This ensures the reliability and usability of the data in various application scenarios, providing strong support for its application. This study offers an effective solution for the efficient synthesis of digital human identity attribute tabular data, significantly contributing to the advancement of related fields.

II. Related Work

The synthesis of digital human identity attributes involves generating fictitious individual identity information through steps such as real data collection, feature extraction, modeling, information generation, and quality assessment. This process allows for the use of synthetic identity information for model training and algorithm testing while preserving privacy. Researchers both domestically and internationally have proposed various innovative methods and frameworks to address different problems and application scenarios. Below are some representative studies:

Choi et al. [5] proposed MedGAN, which combines autoencoders and Generative Adversarial Networks (GAN) to generate multi-label discrete electronic health records (EHR) data. This method effectively handles binary and count variables and demonstrates good performance in practical medical data generation applications. Park et al. [6] introduced TableGAN, utilizing Convolutional Neural Networks (CNN) as both generator and discriminator, and incorporating information loss to improve the quality of generated data. This approach generates tabular data with high authenticity and diversity, proving its effectiveness in various practical scenarios.

For data-sharing scenarios with strict privacy protection requirements, Jordon et al. [7] proposed PATE-GAN, which combines the PATE (Private Aggregation of Teacher Ensembles) framework and introduces perturbations to the teacher discriminator outputs to train the student discriminator, thus achieving differential privacy and generating privacy-protected synthetic data. Xu et al. [8] developed CTGAN, which uses CGAN [9] to address issues of class imbalance and the multimodal distribution of continuous variables. CTGAN excels in multiple real-world datasets, particularly in terms of data similarity and machine learning utility.

Mottini et al. [10] proposed a method based on Cramér GAN [11] and Cross-Net architecture for generating passenger name records (PNR) in the airline industry. This method can handle categorical and numerical features with missing values, and the generated data can be used for business applications such as customer segmentation and nationality prediction of passengers. Koivu et al. [12] introduced actGAN, a minority class oversampling method based on GANs, specifically designed to generate highly imbalanced mixed-type data. In the early stillbirth prediction task, actGAN demonstrated its superiority by significantly improving true positive rates at clinically significant false positive rates.

Zhao et al. [13] proposed CTAB-GAN+, an enhancement over CTGAN, which improves the encoder to better handle mixed-type variables and uses Wasserstein loss with gradient penalty to enhance training stability and effectiveness. CTAB-GAN+ excels in data similarity and analytical utility, making it suitable for generating high-quality synthetic tabular data.

In summary, scholars have conducted extensive research in the field of digital human identity attribute synthesis, proposing various effective methods and frameworks. These studies primarily focus on generating high-quality and diverse tabular data, addressing class imbalance, and ensuring data privacy. They demonstrate the potential and practical effectiveness of Generative Adversarial Networks (GANs) in different application scenarios. However, most studies still face challenges in encoding composite data types or neglect such situations, and none have performed fine-grained evaluations of the quality of synthetic data. This oversight raises the possibility of a lack of social logical consistency.

III. Background Knowledge

To achieve the synthesis system for digital human identity attribute data, this study employs a series of advanced technologies and methods. The following sections will detail the key technologies used in the data generation and processing stages, including CGAN and Vector-Based Sampling,VGM [14] and Association Rule Mining. These technologies play a crucial role in enhancing data generation quality, managing data distribution complexity, and ensuring data consistency.

CGAN introduce conditional variables to the data generation process, allowing the generation of data that adheres to specific attribute requirements. VGM complex data by assuming that the data distribution is a mixture of multiple Gaussian distributions. Association Rule Mining is used to discover relationships between variables in the dataset, ensuring the logical consistency and practical value of the generated data. The integrated use of these technologies enables the construction of an efficient and reliable digital human identity attribute data synthesis system. The following sections provide a brief introduction to these three technologies.

A. CGAN and Vector-Based Sampling

1) CGAN: Conditional Generative Adversarial Networks (CGAN) [9], introduced by Mirza and Osindero, extend the traditional GAN by incorporating conditional variables into both the generator and discriminator, enhancing the model's ability to generate data that meets specific criteria.

In standard GAN, the generator produces data from a random noise vector z, while the discriminator classifies whether the data is real or generated. CGAN, as illustrated in Fig. 1,modifies this by incorporating a conditional variable y. The generator now takes a joint input $[z, y]$ and produces data samples $G(z \mid y)$ corresponding to the condition y. The discriminator evaluates the data based on the joint vector $[x, y]$ and outputs the probability $D(x \mid y)$ that the sample x is real given y. The objective of CGAN is to train the generator to produce samples that the discriminator cannot distinguish from real data, given the condition y. The loss functions for CGAN are adjusted as follows:

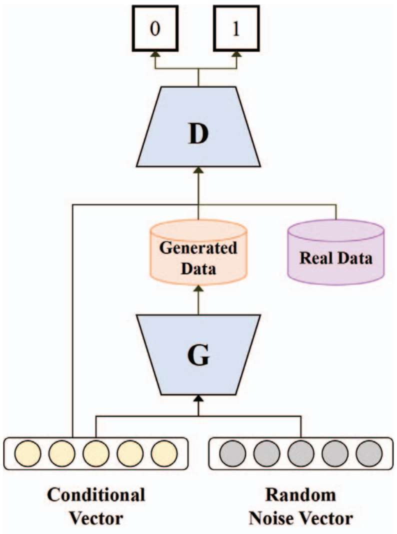

Fig. 1. Schematic Structure of CGAN

$$\begin{aligned} L_{\text{real}}(D, G) &= \mathbb{E}_{x \sim p_{data}(x)}[\log D(x \mid y)] \\ L_{\text{fake}}(D, G) &= \mathbb{E}_{z \sim p_z(z)}\left[\log(1 - D(G(z \mid y)))\right] \\ \min_G \max_D V(D, G) &= L_{\text{real}}(D, G) + L_{\text{fake}}(D, G) \end{aligned}$$

Here, $L_{\text{real}}(D, G)$ represents the expected log probability of real data being correctly classified, and $L_{\text{fake}}(D, G)$ represents the expected log probability of generated data being misclassified as real. The generator aims to create data samples that align with the given condition y, while the discriminator enhances its ability to distinguish between real and synthetic data based on the condition y.

2) Vector-Based Sampling: In CGAN, the sampling training method involves evaluating the generator's output through the discriminator by constructing conditional vectors and sampling data accordingly. For a discrete column D_i with $|D_i|$ values, a conditional vector is created using mask vectors $\mathbf{m}_i$ that indicate specific conditions.

For each sampled discrete column D_i:

- Create N_d zero-filled mask vectors $\mathbf{m}_i$ for all discrete columns.
- Randomly select a column D_{i^*} and determine the condition k^* based on its probability mass function (PMF).
- Set the k^*-th element of $\mathbf{m}_{i^*}$ to 1.
- Construct the conditional vector cond as the concatenation of all mask vectors, including the modified $\mathbf{m}_{i^*}$.

For instance, if $D_1 = \{1, 2, 3, 4\}$ and $D_2 = \{1, 2, 3\}$, a condition $(D_2 = 2)$ is represented by $\mathbf{m}_1 = [0, 0, 0, 0]$ and $\mathbf{m}_2 = [0, 1, 0]$, resulting in cond $= [0, 0, 0, 0, 0, 1, 0]$.

This method ensures that the conditional generator explores all potential values in the discrete columns and helps the discriminator to better estimate the distance between the learned and true conditional distributions.

B. VGM

The VGM is a probabilistic model widely used for data clustering and density estimation. Given that real-world data often exhibit complex, multimodal distributions, VGM effectively captures this complexity by representing the data as a weighted combination of multiple Gaussian distributions, as illustrated in Fig. 2.

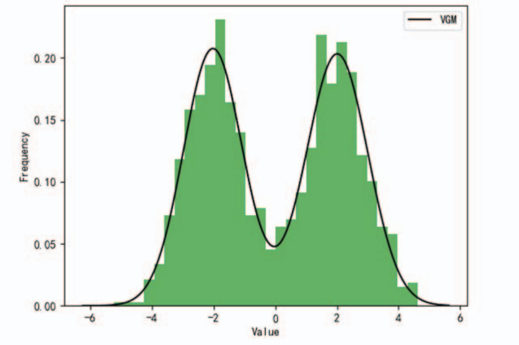

Fig. 2. Example of Distribution Fitting using VGM

In VGM, it is assumed that the data is generated from a mixture of several Gaussian distributions, known as "components." The model can be expressed as a weighted sum of these Gaussian components. Mathematically, the probability density function of VGM is given by:

$$P(X \mid \theta) = \sum_{k=1}^{K} \Pi_k \mathcal{N}(X \mid \mu_k, \Sigma_k)$$

where Π_k represents the weight of the k-th component, μ_k and Σ_k are the mean and covariance matrix of the k-th Gaussian component, respectively, and K is the total number of components.

Parameter estimation for VGM is typically performed using the Expectation-Maximization (EM) algorithm. This iterative method alternates between the Expectation step (E-step) and the Maximization step (M-step). In the E-step, the algorithm computes the expected value of the log-likelihood function given the current estimates of the parameters. In the M-step, the algorithm updates the model parameters to maximize this expected value. The iterative process continues until convergence, refining the parameter estimates.

The EM algorithm updates the parameters as follows:

$$Q(\theta \mid \theta^{(t)}) = \mathbb{E}_{Z|X,\theta^{(t)}}[\log p(X, Z \mid \theta)]$$
$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(t)})$$

where $\theta^{(t)}$ represents the parameters at iteration t, Z denotes the latent variables, and $p(X, Z \mid \theta)$ is the joint probability of the observed data X and the latent variables Z.

Through this iterative optimization process, VGM can effectively handle high-dimensional data, making it suitable for various applications such as clustering, density estimation, and pattern recognition.

C. Association Rule mining

Association analysis, also known as association mining, involves identifying frequent patterns, associations, correlations, or causal structures among items in transaction data, relational data, or other information sources. This technique, which falls under unsupervised learning, is crucial in data mining and is widely used in market basket analysis and recommendation systems. The main goal is to uncover hidden relationships and complex patterns among itemsets in large datasets.

Association mining results in two primary types: frequent itemsets and association rules. Frequent itemsets are collections of items that frequently appear together in transactions, while association rules indicate potential dependencies between items. To evaluate the strength of these associations, two key metrics are used: support and confidence.

- **Support** quantifies the frequency of an itemset or rule across all transactions. It is defined as:

$$\text{Support}(X) = \frac{\text{Number of transactions containing } X}{\text{Total number of transactions}}$$

- **Confidence** measures the conditional probability of itemset Y occurring given that itemset X is present:

$$\text{Confidence}(X \to Y) = \frac{\text{Support}(X \cup Y)}{\text{Support}(X)}$$

Researchers typically set minimum thresholds for support and confidence. An association rule $X \to Y$ is considered strong if it meets or exceeds these thresholds; otherwise, it is deemed weak. The FP-Growth algorithm is used to efficiently discover frequent itemsets and generate association rules in this study.

IV. Method

This study designs and implements a digital human identity attribute data synthesis system based on CGAN. The system design includes the following main modules: data preprocessing, data generation, and data verification. Each module undertakes specific tasks and collectively completes the synthesis of digital human identity attribute data. The following sections will introduce each module in detail.

A. Architecture

The system architecture, as shown in Fig. 3, includes three main modules: data preprocessing, data generation, and data verification. The data preprocessing module is responsible for cleaning, encoding, and normalizing the raw data. The data generation module uses CGAN to generate synthetic data that meets specific conditions. The data verification module employs association rule mining techniques to check the logical consistency of the generated data.

B. Data Preprocessing Module

The main tasks of the data preprocessing module are to clean, encode, and normalize the raw data to facilitate subsequent generation and verification steps. The specific steps are as follows:

- **Data Cleaning:**For a dataset containing digital human identity attribute information, including variables such as ”Gender,” ”Age,” ”Occupation,” and ”Marital Status,” data cleaning is performed to handle corrupted and missing data.
- **Data Encoding:**For discrete variables, one-hot encoding is used. For continuous variables with long-tailed distributions, logarithmic transformation is first applied to reduce their dynamic range. Depending on the complexity of their distribution, either specific mode normalization based on the VGM or general transformation is chosen to accurately capture the complex distribution of continuous variables. For mixed variable types, they are treated as a combination of continuous and discrete variables, with each part being encoded separately.
- **Vector Concatenation:**All encoded data vectors are concatenated according to attribute columns to serve as input for subsequent modules.

C. Data Generation Module

The data generation module is the core component of the system, responsible for generating synthetic data. This module employs CGAN for data generation. By introducing conditional variables, it ensures fair sampling of the real dataset, enabling the generated data to adequately learn the characteristics of rare samples.

- **Generator:**The generator accepts noise vectors and conditional vectors as input and generates high-quality synthetic data through adversarial training.
- **Discriminator** The discriminator distinguishes between real and generated data. Through adversarial training with the generator, it continually improves the quality of the generated data.
- **Conditional Vector:**The conditional vector guides the generator to produce data with specific attributes. By comparing with the distribution of real data, it ensures the authenticity and diversity of the generated data.
- **Fair Sampling Training:**To address the issue of class imbalance in categorical variables, a fair sampling training method is employed. This method adjusts the sampling

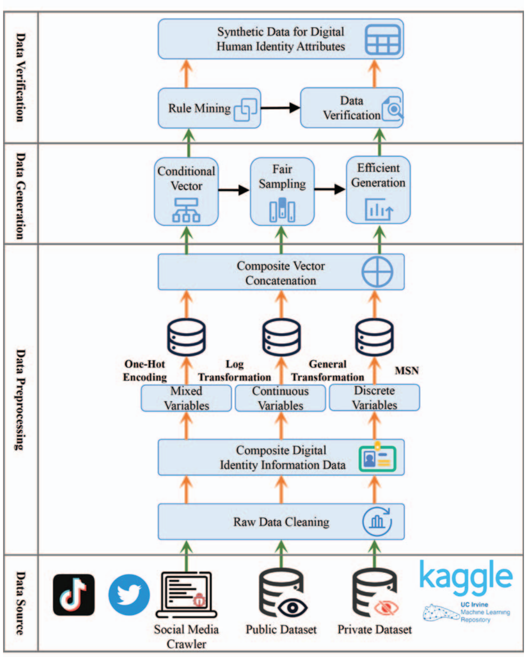

Fig. 3. Architecture of the Identity Attribute Data Synthesis System

strategy of the training data, randomly selecting samples based on the conditional vector. This ensures that less frequent categories receive attention during training, thereby alleviating the problem of category collapse in the synthetic data.

D. Data Verification Module

The data verification module ensures the logical consistency and correctness of the generated data through association rule mining techniques.

- **Association Rule Mining:**Frequent itemsets are mined from real datasets, and absolute association rules with a confidence level of 1 are extracted. This study considers such association dependencies as representations of objective rules in the real world, which should not be altered for the sake of diversity.
- **Verification Process:**After generating the data, the absolute association rules obtained in the previous step are used to verify the synthetic data. This process corrects potential logical errors, ensuring the social logical consistency and correctness of the data.

V. System Testing Experiments

This chapter will conduct ablation experiments on the data verification module within the digital human identity attribute data synthesis system to evaluate its impact on the overall system performance. By comparing systems with and without the data verification module, we will analyze and discuss the results from three aspects: statistical similarity, machine learning utility, and absolute rule retention rate.Additionally, we will visualize the effects of some of the synthetic data to provide an intuitive presentation of the synthesis quality of this system.

A. Experimental Setup

In the proposed system for handling tabular data generation tasks, the following experimental configurations and parameter settings were used: The data generation module was trained on 256×256-sized patches with a batch size of 60 and for 300 epochs. The Adam optimizer was employed with a learning rate of 0.0002 and an exponential decay rate of (0.5, 0.9). The generator utilized a random latent vector with a dimension of 100. In terms of data preprocessing, the data underwent string filling for missing values, corruption checking, and dictionary encoding. Additionally, a gradient penalty term ($\lambda = 10$) was introduced to enhance the stability and convergence of the model.

All experiments were conducted on a server equipped with an NVIDIA RTX 4090 GPU (24GB VRAM), a 16-core vCPU Intel(R) Xeon(R) Gold 6430 processor, and 120GB of RAM. The experiments were implemented using PyTorch 1.11.0 and Python 3.7, ensuring the reliability and reproducibility of the experimental results.

B. Evaluation Metrics

The evaluation metrics primarily include statistical similarity differences, machine learning utility differences, and absolute rule retention rate.

*a) **Statistical Similarity Differences**:* Jensen-Shannon Divergence (JSD) and Wasserstein Distance (WD) are used to measure the distributional similarity between the generated data and real data. The difference in feature correlations (Diff. Corr.) evaluates the preservation of relationships between features. Lower values of JSD, WD, and Diff. Corr. indicate higher distributional similarity between the generated and real data.

*b) **Machine Learning Utility Differences**:* This metric assesses the practical utility of the generated data in training machine learning models. By training classification models with the generated data and evaluating the models' accuracy, precision, recall, and F1 score on a test set, we compare these results to models trained on real data. The differences in these evaluation metrics are then calculated. Lower utility difference values suggest that the generated data better captures the distributional characteristics of the real data, providing higher practical value for training machine learning models.

*c) **Absolute Rule Retention Rate**:* Using association rule mining techniques, high-confidence association rules are extracted from the real data, and their retention rate in the generated data is calculated. A higher rule retention rate indicates that the generated data is more reliable in terms of logical consistency and practical application.

C. Experimental Results and Analysis

To evaluate the authenticity of the proposed system in tabular data generation scenarios, experiments were conducted using the Adult dataset [15]. The experimental results are

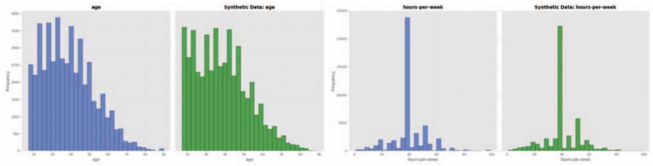

Fig. 4. Comparison of Linear Distributions

summarized in Table I, which presents the results of the ablation study based on several evaluation metrics.

TABLE I
EXPERIMENT RESULTS

Metric	w/ Data Verification	w/o Data Verification
JSD	0.043	0.048
WD	0.004	0.004
Diff.Corr.	0.273	0.309
Acc	1.128%	1.218%

In addition to the quantitative results, Table II provides an illustrative example of the synthetic data generated by the system, showcasing a subset of the vector columns. Fig. 4 compares the linear distribution of synthetic data against real data, offering a visual representation of how well the synthetic data aligns with the actual data distribution.

TABLE II
SYNTHETIC DATA EXAMPLES

Age	Workclass	Fnlwgt	Education	Income	...
49	Private	193366	HS-grad	≤50K	...
23	Local-gov	190709	Assoc-acdm	≤50K	...
56	Local-gov	216851	Bachelors	≥50K	...
48	Self-emp-not-inc	83311	Bachelors	≤50K	...

Experimental results demonstrate that our system excels in both statistical similarity differences and machine learning utility, indicating that it effectively learns the characteristics of real datasets, resulting in synthetic data with high authenticity and accuracy. Additionally, the introduction of the data verification module significantly enhances the absolute rule retention rate. This suggests that the system with the verification module can fully capture the absolute associative relationships from real datasets, effectively maintaining the logical consistency of the synthetic data and increasing its applicability across various fields.

VI. CONCLUSION

This study proposes a digital human identity attribute data synthesis system based on CGAN. Through three modules—data preprocessing, data generation, and data verification—the system achieves the generation of high-quality data. Experimental results indicate that our system performs well in terms of statistical similarity and machine learning utility. Furthermore, it demonstrates superior performance in terms of absolute rule retention rate compared to systems without the data verification module, validating the importance of the data verification module in enhancing the quality and consistency of the generated data. The system presented in this paper provides strong support for the efficient synthesis of digital human identity attribute data, holding significant importance for the advancement of related fields. Future work can focus on further optimizing the data verification module and exploring its application in larger-scale datasets and more diverse scenarios.

REFERENCES

[1] Chen R J, Lu M Y, Chen T Y, et al. Synthetic data in machine learning for medicine and healthcare[J]. Nature Biomedical Engineering, 2021, 5(6): 493-497.

[2] Abbasi A, Albrecht C, Vance A, et al. Metafraud: a meta-learning framework for detecting financial fraud[J]. Mis Quarterly, 2012: 1293-1327.

[3] El Emam K, Arbuckle L. Anonymizing health data: case studies and methods to get you started[M]. " O'Reilly Media, Inc.", 2013.

[4] Borisov V, Leemann T, Seßler K, et al. Deep neural networks and tabular data: A survey[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022.

[5] Choi E, Biswal S, Malin B, et al. Generating multi-label discrete patient records using generative adversarial networks[C]//Machine learning for healthcare conference. PMLR, 2017: 286-305.

[6] Park N, Mohammadi M, Gorde K, et al. Data synthesis based on generative adversarial networks[J]. arXiv preprint arXiv:1806.03384, 2018.

[7] Jordon J, Yoon J, Van Der Schaar M. PATE-GAN: Generating synthetic data with differential privacy guarantees[C]//International conference on learning representations. 2018.

[8] Xu L, Skoularidou M, Cuesta-Infante A, et al. Modeling tabular data using conditional gan[J]. Advances in neural information processing systems, 2019, 32.

[9] Mirza M, Osindero S. Conditional generative adversarial nets[J]. arXiv preprint arXiv:1411.1784, 2014.

[10] Mottini A, Lheritier A, Acuna-Agost R. Airline passenger name record generation using generative adversarial networks[J]. arXiv preprint arXiv:1807.06657, 2018.

[11] Bellemare M G, Danihelka I, Dabney W, et al. The cramer distance as a solution to biased wasserstein gradients[J]. arXiv preprint arXiv:1705.10743, 2017.

[12] Koivu A, Sairanen M, Airola A, et al. Synthetic minority oversampling of vital statistics data with generative adversarial networks[J]. Journal of the American Medical Informatics Association, 2020, 27(11): 1667-1674.

[13] Zhao Z, Kunar A, Birke R, et al. Ctab-gan+: Enhancing tabular data synthesis[J]. Frontiers in big Data, 2024, 6: 1296508.

[14] Christopher M Bishop. Pattern recognition and machine learning. springer, 2006.

[15] D. Dua and C. Graff, "UCI Machine Learning Repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml. [Accessed: 15-Sep-2024].