Fair and Privacy-preserving Synthetic Data Generation via Clustering-based Variational Autoencoder and Adversarially Debiased Wasserstein Generative Adversarial Networks with Gradient Penalty

Malek Adouani[1] ✉ and Zaineb Chelly Dagdia[1]

[1]Université Paris-Saclay, UVSQ, DAVID, France
malek.adouani@uvsq.fr, zaineb.chelly-dagdia@uvsq.fr

Abstract. The increasing reliance on machine learning in sensitive domains, such as healthcare, has amplified concerns about bias and privacy in data-driven decision-making. While fairness-aware generative models aim to mitigate bias, they often depend on labeled data, limiting their applicability in unsupervised settings. Conversely, differentially private generative models ensure privacy but may still encode hidden biases. Existing methods fail to jointly optimize fairness and privacy without explicit supervision. To address this gap, we propose a hybrid generative framework that integrates clustering-based Variational Autoencoder (VAE) with Wasserstein Generative Adversarial Networks with Gradient Penalty (WGAN-GP) to generate fair and privacy-preserving synthetic data. The VAE structures latent representations under zero-Concentrated Differential Privacy (zCDP) while incorporating K-Means clustering directly in the latent space. The clustering serves as a factor to influence the generative process into producing samples that resemble real data in unsupervised settings. These structured representations along with cluster labels then guide WGAN-GP's generator toward sample generation and enhance adversarial debiasing through the Fairness Critic, which penalizes correlations between synthetic data and sensitive attributes to ensure fairness. By integrating clustering-based VAEs with WGAN-GP, our framework enforces fairness while maintaining strong privacy guarantees. Experimental results demonstrate that it outperforms existing generative models by effectively reducing bias, preserving privacy, and ensuring high data utility across multiple fairness and privacy metrics.

Keywords: Generative Adversarial Networks · Unsupervised learning · Bias mitigation · Privacy preservation.

1 Introduction

Machine learning models are increasingly deployed in critical applications such as healthcare and finance, where biased and privacy-compromising decisions can

have serious societal consequences. Models trained on biased datasets risk reinforcing societal disparities, leading to discriminatory outcomes that disproportionately affect marginalized groups [10]. While fairness-aware machine learning methods attempt to mitigate bias, many rely on labeled data, which are often scarce, costly to obtain, and may not fully capture the diversity of real-world populations [15]. When labels are incomplete or unavailable, fairness assessment becomes more challenging, as bias can manifest in hidden ways within latent representations and sampling distributions [12]. At the same time, privacy concerns in sensitive domains necessitate mechanisms that protect individual data while enabling meaningful analysis. Differential Privacy [9] has emerged as a strong privacy-preserving solution, ensuring that synthetic datasets do not reveal information about specific individuals. However, privacy-preserving generative models often fail to address hidden biases in data, as differential privacy constraints can obscure fairness-related information rather than eliminate it [11]. This tradeoff between fairness and privacy presents a fundamental challenge in generative modeling, particularly when working with unlabeled data. Generative models, especially Generative Adversarial Networks (GANs), offer a promising approach to producing realistic synthetic data that preserves statistical properties of the original dataset. Fairness-aware GANs introduce debiasing constraints [10], but their reliance on labeled data limits their utility in unsupervised scenarios. Conversely, differentially private GANs prioritize privacy but often fail to mitigate bias, potentially encoding and perpetuating hidden disparities in generated samples [11]. This raises a crucial open problem: How can we jointly enforce fairness and privacy in generative modeling without requiring labeled data?
To address this challenge, we propose a hybrid generative framework that integrates fairness-aware latent space structuring with privacy-preserving mechanisms, enabling the generation of fair and privacy-preserving synthetic data in an unsupervised setting. Our approach combines:

- Clustering-based Variational Autoencoder (VAE) with zero-Concentrated Differential Privacy (zCDP): The clustering based VAE plays a pivotal role in structuring the latent space while preserving privacy. By incorporating K-Means clustering directly into the latent representation, the model enforces the grouping of similar data points, ensuring that downstream generative processes capture the structured distributions of real data in unsupervised settings. The cluster labels serve as guiding signals for the subsequent adversarial training phase, enabling controlled sample generation that aligns with the underlying data structure. The variational inference framework further enhances the model's ability to learn meaningful data representations, improving the quality of generated representations [21]. To ensure strong privacy guarantees, we enforce zCDP, which bounds the moments of privacy loss random variable rather than imposing a fixed limit, making it a more refined and mathematically rigorous privacy mechanism. This approach avoids the infinite loss scenarios associated with traditional (ϵ, δ)-DP (Differential Privacy) while offering tighter privacy guarantees [9]. By structuring the latent space through clustering-based enforcement and mathematically rigorous privacy constraints, our VAE component provides a strong foundation for generating high-quality, fair, and privacy-compliant synthetic data.

- Adversarially Debiased Wasserstein GAN with Gradient Penaty (WGAN-GP): While VAE ensures fairness in latent representations and enforces privacy, it does not directly control fairness at the synthetic data level. To achieve this, we deploy WGAN-GP which stabilizes training and mitigates mode collapse challenges that often arises in GANs trained on biased data [10]. Crucially, the cluster-aware latent representations obtained from the VAE serve as conditional inputs to the GAN, ensuring that the generator produces samples that align with the structured distributions discovered in an unsupervised manner. Additionally, we introduce an adversarial debiasing mechanism via the fairness critic, which penalizes unwanted correlations between generated data and sensitive attributes. This technique has been shown to be effective in reducing bias in generative models by directly enforcing fairness constraints through adversarial optimization [7]. The training process ensures a tradeoff between data realism and fairness, where the clustering-based VAE conditions the generator while the fairness critic refines the final output to achieve unbiased data generation.

By explicitly addressing bias in the generative process and incorporating zCDP, our framework named “Clust-VAE-WGAN-GP” bridges the gap between fairness-aware and privacy-preserving generative modeling in unsupervised settings. This advancement is valuable for applications where explicit labels are unavailable, enabling more ethical and reliable synthetic data generation.

2 Related Work

Ensuring fairness in synthetic data generation remains a critical challenge in machine learning where models trained on biased data can exacerbate societal inequalities. Bias arises from sources such as covariate shift, selection bias, and class imbalance, leading to discriminatory outcomes that disproportionately affect certain demographic groups [1]. A major challenge lies in labeled data, as machine learning models often assume datasets are representative of the population—an assumption that rarely holds in real-world applications [15]. Generative models, such as VAEs and GANs, offer promising solutions for bias mitigation by learning complex data distributions. However, fairness-aware generative modeling remains challenging, particularly in unsupervised settings where bias can propagate through latent representations and sampling distributions [3]. FairGAN [7] and its improved version FairGAN+ [8] introduced fairness constraints to enforce statistical parity across sensitive attributes, ensuring that generated samples do not disproportionately favor specific demographic groups. TabFairGAN [2] extends these techniques to tabular data, leveraging Wasserstein GAN (WGAN) to create demographically balanced datasets by adjusting sample distributions. Similarly, conditional GANs (cGANs) [16] aim to reduce bias by conditioning on fairness-related constraints, thereby promoting balanced representations in generated samples. However, these methods rely on explicit class labels, making them unsuitable for unsupervised generative modeling, where such labels are unavailable. Furthermore, fairness evaluation in generative models is inherently challenging, as even highly accurate sensitive attribute classifiers introduce measurement errors that can distort fairness assessments [3]. Given the limitations of labeled data, semi-supervised learning has emerged as a promising approach to enhance fairness by leveraging unlabeled data. [20], for instance, demonstrated its effectiveness in mitigating class imbalance in fault detection diagnosis, while in medical AI, it has been shown to improve fairness by capturing broader data distributions and reducing bias in predictive models [6]. By compensating for biased labeled datasets, semi-supervised methods lead to more generalizable and fairer models [15]. Another emerging approach is Positive-Unlabeled Learning, where models learn from datasets containing only positive samples and unlabeled data [14]. Observer-GAN [13] introduced an observer network to generate pseudo-negative samples, allowing the model to differentiate between groups without explicit labels, thereby improving fairness and generalization in synthetic data generation. In parallel to fairness-aware methods, researchers have focused on privacy-preserving generative models to ensure that synthetic data does not expose sensitive individual information. Differentially Private GANs (DPGANs) [4] and PATEGAN [19] integrate differential privacy (DP) mechanisms to prevent data leakage by injecting noise into the model's gradients or outputs. Meanwhile, RDPCGAN [5] leverages Rényi Differential Privacy (RDP) to achieve even stronger privacy guarantees while preserving data utility. However, a key limitation of most differentially private GANs is their exclusive focus on privacy, often neglecting fairness constraints. As a result, bias can persist in privacy-preserving synthetic datasets, potentially exacerbating disparities if left unaddressed. Despite progress in fairness-aware generative modeling, unsupervised approaches remain underexplored. Existing methods either require labeled data to enforce fairness or focus solely on privacy, overlooking bias mitigation. This gap highlights the need for a hybrid generative model that ensures fairness without explicit class labels while preserving privacy, enabling ethical and reliable synthetic data generation in sensitive domains.

3 Proposed Method

This section introduces our Clust-VAE-WGAN-GP hybrid generative framework, which integrates a Cluster-Based VAE with zCDP and a WGAN-GP enhanced by a Fairness Critic.

3.1 Overall Architecture

As illustrated in Figure 1, the model operates in two interconnected stages. In the first stage, the Cluster-Based VAE encodes input data into a continuous latent space, where K-Means clustering is applied to structure the latent representations. This clustering mechanism generates cluster embeddings that influence the generative process, guiding it to produce synthetic samples that closely resemble real ones in an unsupervised setting. This is achieved by conditioning the generator in the second stage on the clustering information, ensuring that the generated data aligns with the learned latent structure. To enforce privacy, zCDP regulates information leakage through Gaussian noise injection. WGAN-GP utilizes the resulting clustered latent encodings (i.e., continuous private samples), along with their assigned cluster labels outputted from the Cluster-Based VAE with zCDP. The generator is explicitly conditioned on these cluster labels, ensuring that generated samples preserve the latent structure learned in the unsupervised setting. This conditioning mechanism helps maintain distributional consistency across clusters, enhancing data realism. The Fairness Critic further enforces fairness by evaluating correlations between generated samples and predefined sensitive attributes, guiding the generator to reduce statistical dependence on these attributes via adversarial debiasing. The discriminator randomly receives samples from the generator and the continuous private samples to determine and tries to distinguish real from fake samples. Its output is regulated through Wasserstein loss and Gradient penalty for smoother update to the generator. By integrating structured latent encodings from the VAE with adversarial fairness constraints in WGAN-GP, our model generates high-fidelity, privacy-preserving, and bias-free synthetic data, making it suitable for fairness-aware applications in privacy-sensitive domains. Algorithm 1 presents the Clust-VAE-WGAN-GP functioning.

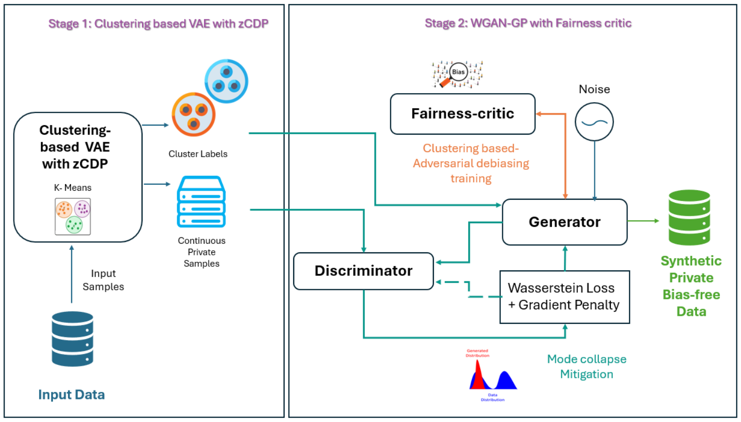

Stage 1: Clustering based VAE with zCDP
Clustering-based VAE with zCDP
K- Means
Cluster Labels
Continuous Private Samples
Input Samples
Input Data
Stage 2: WGAN-GP with Fairness critic
Bias
Fairness-critic
Noise
Clustering based-Adversarial debiasing training
Generator
Discriminator
Wasserstein Loss + Gradient Penalty
Synthetic Private Bias-free Data
Generated Distribution
Data Distribution
Mode collapse Mitigation

Fig. 1: Overall architecture of Clust-VAE-WGAN-GP.

3.2 Detailed Architecture

Let $\mathcal{X}$ be the input data space, where each sample $\mathbf{x} \in \mathcal{X}$ is drawn from a biased dataset. The objective is to learn a synthetic data generator that satisfies the following conditions: (i) The generated data preserves privacy by adhering to differential privacy constraints, preventing individual re-identification. (ii)

The generative process mitigates bias by minimizing correlations between generated data and predefined sensitive attributes, ensuring fair and representative sample distributions using fairness critic. (iii) The model learns cluster-aware latent representations and promotes data realism, making it suitable for unsupervised learning tasks. To ensure these conditions, the components of Clust-VAE-WGAN-GP are structured as follows.

Clustering-Based VAE with zCDP Figure 2 illustrates the VAE component, which integrates probabilistic encoding, clustering, and zCDP to generate structured, private latent encodings. Given an input sample x, the encoder maps it to a latent representation z by approximating the posterior distribution [21]:

$$q_\phi(z|x) = \mathcal{N}(\mu_\phi(x), \sigma^2_\phi(x)) \tag{1}$$

where $\mu_\phi(x)$ and $\sigma_\phi(x)$ are the mean and standard deviation parameters, respectively, learned by the encoder network. The reparameterization trick ensures continuous and differentiable latent representations. The training objective minimizes the KL-divergence (KL-loss) between $q_\phi(z|x)$ and the prior $p(z)$, structuring the latent space to align with the original data distribution. $p(z)$ is used to generate deterministic points $\hat{x}$ using KL-loss regularization. The decoder reconstructs the input $\hat{x}$ into $\hat{x}_c$, while minimizing the reconstruction loss, preserving essential data properties. To resemble real data in the unsupervised setting, K cluster centroids $c_{k}{}_{k=1}^{K}$ are initialized in the latent space $\hat{x}$, and proximity constraints are enforced via K-Means clustering. These will be next given as an input to the WGAN-GP. For privacy preservation, zCDP is integrated into the VAE by adding Gaussian noise to gradients during backpropagation, mitigating the risk of sensitive information leakage. The privacy loss ρ is given by the following equation, where q is the sampling probability, σ is the noise multiplier, and T is the total number of training iterations [9]:

$$\rho = \frac{q^2 \sigma^{-2} T}{2} \tag{2}$$

The structured latent encodings $\hat{x}_c$ and their corresponding cluster labels c_k form the basis for training the second stage of the model, ensuring that the WGAN-GP component adheres to fairness and privacy constraints.

WGAN-GP with Adversarial Fairness Critic Figure 3 presents the second stage of the framework, where a WGAN-GP learns to generate structured synthetic data while enforcing fairness constraints.

The latent encodings $\bar{x}_c$ and cluster labels generated by the VAE serve as input to the WGAN-GP component, where the labels constitute a condition for the generator during GAN training. The generator G_θ receives as input a combination of structured latent noise $\bar{z} \sim p(z)$ and cluster label embeddings c_k, producing synthetic samples $\bar{x}_f$ that tend to resemble real data and their assignment to the same cluster. The discriminator D_ψ is trained to differentiate

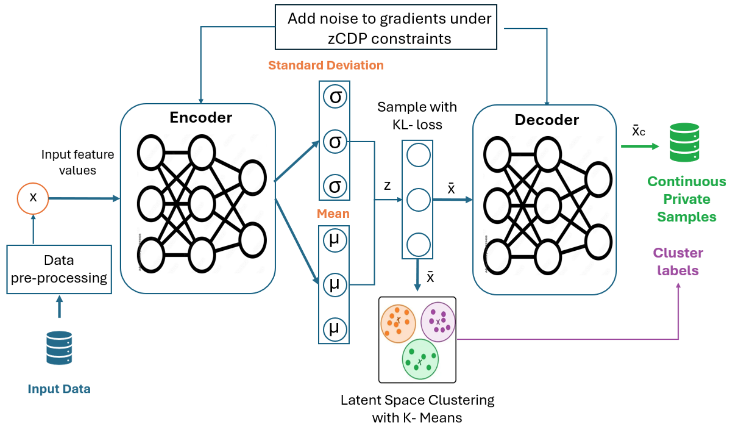


Fig. 2: Overview of Clustering-Based VAE with zCDP mechanism.

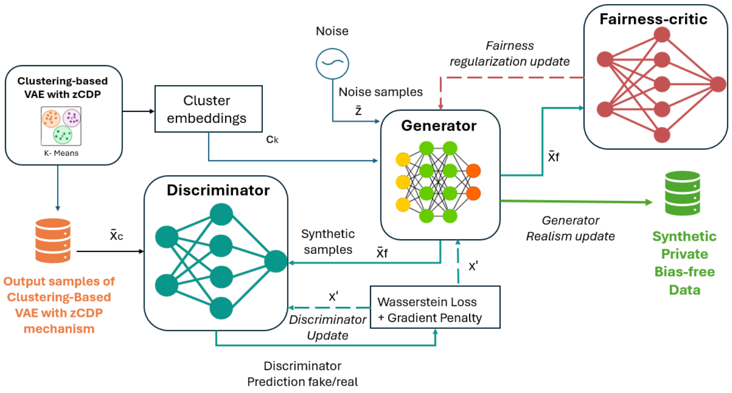

Noise
Fairness-critic
Fairness regularization update
Clustering-based VAE with zCDP
K- Means
Cluster embeddings
Noise samples
z̄
Ck
Generator
x̄f
Discriminator
x̄c
Synthetic samples
x̄f
x'
Generator Realism update
Synthetic Private Bias-free Data
Output samples of Clustering-Based VAE with zCDP mechanism
x'
Discriminator Update
Wasserstein Loss + Gradient Penalty
Discriminator Prediction fake/real

Fig. 3: Overview of WGAN-GP with a Fairness Critic.

between real samples $\bar{x}_c$ and synthetic samples $\bar{x}_f$ while optimizing the revised WGAN-GP loss:

$$\mathcal{L}_{\text{WGAN-GP}} = \mathbb{E}[D_\psi(\bar{x}_f)] - \mathbb{E}[D_\psi(\bar{x}_c)] + \lambda\mathbb{E}[(\|\nabla_{x'} D_\psi(x')\| - 1)^2] \tag{3}$$

where x' represents interpolated samples between real and generated data, stabilizing training. The fairness critic C_φ enforces fairness by penalizing dependence to sensitive attributes s from generated samples. Since the generator is conditioned on structured cluster labels, the Fairness Critic can detect and penalize correlations between synthetic samples and sensitive attributes, minimizing bias. The generator is updated to reduce statistical dependence on these attributes via the fairness loss:

$$\mathcal{L}\text{fair} = \mathbb{E}[|C_\varphi(G_\theta(\bar{z}, c_k)) - s|^2] \tag{4}$$

To ensure trade-off between data realism and fairness constraints, the synthetic samples $\bar{x}_f$ are assigned cluster labels based on precomputed cluster embeddings. These embeddings, obtained from the Clustering-based VAE, remain fixed during WGAN-GP training. This conditioning reinforces both bias mitigation and structured generation while ensuring consistency in cluster-based representation.

Algorithm 1 The Clust-VAE-WGAN-GP algorithm

```
Require: Dataset $\mathcal{X}$, learning rates $\eta_1, \eta_2, \eta_3$, batch size $B$, number of clusters $K$,
    privacy noise $\sigma$, Wasserstein loss weight $\lambda$
Initialize: Clustering-based VAE $V_\phi$, WGAN-GP $(G_\theta, D_\psi)$, Fairness Critic $C_\varphi$
Compute zCDP privacy parameters $(\epsilon, \delta)$ using Gaussian mechanism
Pre-train VAE: Encode input $\mathcal{X}$ into latent space, compute mean and variance,
    and apply KL loss
Initialize cluster centroids using K-Means on VAE latent representations
for each training epoch do
    for each batch $\mathcal{B} \subset \mathcal{X}$ do
        Encode $\mathcal{B}$ using VAE, inject Gaussian noise $\mathcal{N}(0, \sigma^2)$, sample latent representations
        Assign cluster labels based on precomputed cluster embeddings
        Train WGAN-GP:
            Sample noise $\mathbf{z}$ and cluster embeddings to generate synthetic data
            Compute Wasserstein loss with gradient penalty
            Update generator $G_\theta$ and discriminator $D_\psi$
        Train Fairness Critic $C_\varphi$:
            Predict sensitive attributes from generated samples
            Compute fairness loss and update $G_\theta$ to remove bias
    end for
end for
Return: Trained models $V_\phi$, $G_\theta$, $D_\psi$, and $C_\varphi$
```

4 Experimental setup

We investigate the following Research Questions (RQ) using the datasets detailed in Table 1: (1) RQ1: How can we generate synthetic data that closely resembles real data while maintaining high fidelity? (RQ2) : How can we balance the trade-off between privacy and utility in synthetic data generation? (3) RQ3: How can synthetic data generation effectively mitigate bias? (4) RQ4: How can we guarantee the trade-off between privacy and fairness? (RQ5): How can we ensure that synthetic data preserves privacy?

For hyperparameter tuning, the VAE with clustering uses 4 layers in both the encoder and decoder, while the discriminator and fairness critic use 3 layers, and

Table 1: Overview of Benchmark Datasets Characteristics.

Dataset	# Instances	# Features	Labels	# Numerical Features	# Categorical Features
HIV	8916	22	Unlabeled	3	19
Corporate Stress	3000	45	Binary	37	8
Obesity	2111	17	Binary	8	9
Heart Failure	299	13	Binary	12	1
Diabetes Dataset	253680	22	Binary	21	1
Pediatric	782	58	Binary	19	39

the generator uses 4 layers. The LeakyReLU activation function is applied. We chose a batch size of 100. The noise multiplier ranges from 0.1 to 0.5. The learning rate is set to 0.0001. Sensitive attributes, such as Age, Gender, and Ethnicity, are manually identified for each dataset. A privacy accountant is used to convert ρ, the privacy loss parameter of zCDP, into (ϵ, δ)-DP for a fair comparison with baseline methods using (ϵ, δ)-DP. The number of clusters for K-Means is varied from 10 to 25. Kindly refer to our method's source code at the following link: Unsupervised Cluster-based VAE WGAN GP.

To assess *synthetic data realism*, the quality of synthetic data and its similarity to real data, we used the following metrics [16]:

- **Maximum Mean Discrepancy (MMD)**: Measures the distributional distance between real and synthetic data, with lower values indicating higher realism.
- **Kolmogorov-Smirnov (KS) Test**: The KS test assesses whether real and synthetic data follow the same distribution by measuring the maximum difference between their cumulative distribution functions. A lower KS value indicates stronger similarity. If the p-value is greater than 0.05, the difference is not statistically significant, suggesting that the synthetic data may follow the same distribution as the real data.
- **Wasserstein Distance (WD)**: Measures the optimal transport cost required to match the synthetic data distribution to the real data distribution, where lower values indicate better alignment.
- **Dimension-Wise Probability (DWP) Score**: Evaluates the per-feature distribution similarity between real and synthetic datasets. A score closer to 1 indicates greater similarity, meaning the synthetic data closely follows the real data's feature distributions.
- **Alpha-Precision**: Measures how many synthetic samples lie within the support of the real data distributions. A value closer to 1 indicates better utility.
- **Beta-Recall**: Measures how much of the real data distribution is covered by the synthetic data. A value closer to 1 indicates better utility.

Since explicit labels are unavailable, *fairness* is evaluated based on how sensitive attributes are represented within clusters in the generated data:

- **Statistical Parity in Clusters (SP)** [18]: Ensures that sensitive attributes do not influence cluster assignments, promoting fairness.
- **Mutual Information (MI) Between Clusters and Sensitive Attributes [18]**: Quantifies the dependency between synthetic cluster assignments and sensitive attributes. Lower MI values indicate better fairness.
- **Cluster Quality and Bias Impact** [17]:
    - **Silhouette Score (SS)**: Measures how well data points fit within their assigned clusters, with values closer to 1 indicating well-defined, unbiased clusters.
    - **Davies-Bouldin Index (DBI)**: Evaluates intra-cluster cohesion and inter-cluster separation, where lower values indicate better clustering performance.

To evaluate the privacy risks associated with synthetic data, we utilized the following metrics [16]:

- **Epsilon Identifiability Risk:** Quantifies the probability that an individual record can be uniquely identified in synthetic data. A higher epsilon risk suggests potential re-identification threats.
- **Nearest Neighbour Distance Ratio (NNDR):** Evaluates how distinguishable real data points are from synthetic ones based on nearest-neighbour distances. A lower NNDR means better privacy preservation.

We compared our Clust-VAE-WGAN-GP method with the following baseline approaches to assess its potential in mitigating bias and preserving privacy.

- **FairGAN** [7]: Introduces fairness constraints to enforce statistical parity across sensitive attributes in generated data.
- **TabFairGAN** [2]: Extends FairGAN to tabular data using Wasserstein GAN (WGAN) for demographically balanced datasets.
- **DPGAN** [4]: Ensures privacy in synthetic data generation by incorporating differential privacy mechanisms.
- **RDPCGAN** [5]: Utilizes Rényi Differential Privacy (RDP) to achieve strong privacy guarantees while maintaining data utility.

5 Results and discussion

5.1 Data realism evaluation

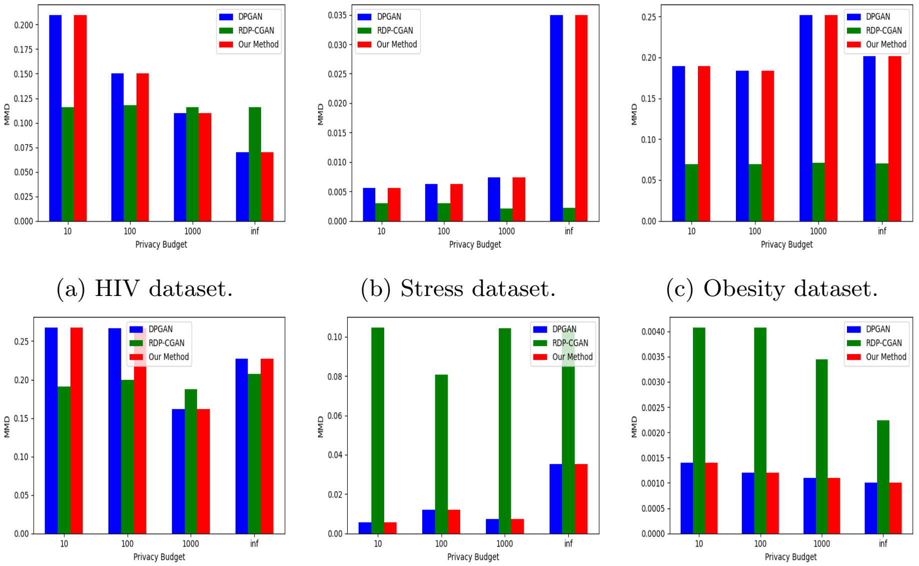
Evaluation under no privacy constraints To address RQ1, Table 2 compares the utility of synthetic data generated by our method against baseline approaches. The results show that our model consistently outperforms baseline methods across most datasets and remains highly competitive with TabFairGAN. Specifically, it achieves the best performance in four out of six realism metrics for the Heart and Pediatric datasets. In three other datasets (HIV, Stress, and Diabetes), our method shares top-performing metrics with TabFairGAN. For instance, in the Pediatric dataset, our method achieved the lowest MMD and

WD values (0.0010, 0.238), outperforming TabFairGAN (0.1890, 0.25), FairGAN (0.2011, 0.49), DPGAN (0.2154, 0.55), and RDP-CGAN (0.7801, 0.30). It also attained the highest DWP score (0.5977), α-precision (0.741), and β-recall (0.712), clearly surpassing all baselines. In the Heart dataset, our method leads across all six metrics, with a DWP score of 0.675, α-precision of 0.715, and β-recall of 0.712. In the HIV dataset, it achieved the best DWP score (0.63) and β-recall (0.749), while remaining competitive on the other metrics. In Stress and Diabetes, it shared top performance: achieving the highest DWP scores (0.501, 0.5018) and the highest α-precision and β-recall in both datasets (Stress: 0.770, 0.509; Diabetes: 0.678, 0.888). The superior performance of our method stems from the integration of variational autoencoders with variational inference, enabling it to effectively capture feature correlations. Additionally, the novel conditioning mechanism introduced by K-Means clustering within the latent space guides the generative process, ensuring that synthetic samples closely resemble real data in an unsupervised setting. However, TabFairGAN remains a strong competitor, sharing several top-performing metrics with our method in the HIV, Stress, and Diabetes datasets and outperforming it in the Obesity dataset. This advantage can be attributed to its generator architecture, which incorporates ReLU activation for numerical attributes and Gumbel-softmax for categorical features, enhancing its ability to generate mixed-type data—particularly beneficial for the Obesity dataset.

Evaluation under privacy constraints To answer RQ2, Figure 4 illustrates MMD trends under varying privacy budgets. Our method, in almost all datasets, consistently achieves lower MMD values at higher privacy budgets (e.g., 1000 and ∞), demonstrating a strong tradeoff between privacy and utility. However, at lower privacy budgets, RDP-CGAN shows competitive performance with lower MMD values. This can be attributed to the stronger privacy bound imposed by zCDP in our approach, which can introduce more noise at lower privacy budgets. Meanwhile, standard DP mechanisms, used in DP-GAN are vulnerable to exponential privacy loss accumulation causing decreased utility. Additionally, we observe instability in MMD values at higher privacy budgets for all models in the case of the Stress and Diabetes datasets, where MMD values do not exhibit a descending pattern as the privacy budget increases. This instability arises from dataset characteristics: the Diabetes dataset's large size and moderate feature count make it difficult to balance privacy and utility, as reduced noise can cause the generator to focus on irrelevant patterns, leading to less reliable results. The Stress dataset has a moderate size and includes a mix of binary and categorical workplace stress indicators, contributing to the observed instability in MMD values. This combination of mixed data types and weakly structured patterns makes it challenging for the generator to capture meaningful relationships. This results to the vulnerability of the generator to overfitting noise rather than learning useful representations increasing instability.

Table 2: Data Realism Metrics Across Methods under no privacy constraints.

Dataset	**Method**	**MMD**	**KS-test**	**WD**	**DWPscore**	**α-precision**	**β-recall**
HIV	Our Method	0.07	**0.09**	0.43	**0.63**	0.567	**0.749**
	TabFairGAN	**0.0398**	0.07	**0.057**	0.4418	**0.692**	0.690
	FairGAN	0.1461	0.03	0.57	0.4294	0.680	0.680
	DPGAN	0.1452	3.50e-46	0.58	0.3481	0.600	0.650
	RDP-CGAN	0.9646	1.65e-6	0.30	0.3830	0.510	0.580
Stress	Our Method	0.035	**0.28**	0.28	**0.501**	**0.770**	0.509
	TabFairGAN	**0.0011**	0.51	**0.036**	0.4571	0.690	**0.690**
	FairGAN	0.0030	2.14e-68	0.480	0.3124	0.650	0.670
	DPGAN	0.0028	2.17e-56	0.51	0.3578	0.610	0.630
	RDP-CGAN	0.8488	3.380e-2	0.35	0.2304	0.530	0.550
Obesity	Our Method	0.2013	**0.05**	0.56	0.43	0.439	0.487
	TabFairGAN	**0.0213**	0.087	**0.083**	**0.4948**	**0.697**	**0.581**
	FairGAN	0.5609	1.19e-86	0.97	0.2457	0.600	0.590
	DPGAN	0.0876	1.237e-9	0.97	0.3159	0.570	0.610
	RDP-CGAN	1.0431	5.41e-2	0.80	0.1551	0.480	0.520
Heart	Our Method	0.227	**0.1701**	**0.227**	**0.675**	**0.715**	**0.712**
	TabFairGAN	**0.2674**	0.07	0.19	0.3194	0.690	0.690
	FairGAN	0.2300	6.55e-12	0.37	0.2760	0.680	0.690
	DPGAN	0.2387	9.34e-5	0.531	0.3576	0.630	0.660
	RDP-CGAN	0.7563	7.14e-1	0.34	0.1941	0.500	0.550
Diabetes	Our Method	**0.0352**	0.035	0.285	**0.5018**	**0.678**	**0.888**
	TabFairGAN	0.1782	**0.09**	**0.22**	0.3750	0.612	0.690
	FairGAN	0.1942	5.87e-9	0.51	0.3014	0.620	0.650
	DPGAN	0.2083	3.22e-7	0.57	0.2998	0.580	0.600
	RDP-CGAN	0.7999	7.22e-4	0.28	0.2554	0.490	0.500
Pediatric	Our Method	**0.0010**	0.03	**0.238**	**0.5977**	**0.741**	**0.712**
	TabFairGAN	0.1890	**0.05**	0.25	0.3950	0.690	0.690
	FairGAN	0.2011	3.21e-8	0.49	0.3198	0.670	0.690
	DPGAN	0.2154	2.98e-6	0.55	0.3120	0.630	0.640
	RDP-CGAN	0.7801	6.11e-3	0.30	0.2679	0.520	0.560



(a) HIV dataset.

(b) Stress dataset.

(c) Obesity dataset.

(d) Heart Failure dataset.

(e) Diabetes dataset.

(f) Pediatric dataset.

Fig. 4: Mean Maximum Discrepancy (MMD) comparison across different methods (DPGAN, RDP-CGAN, and our approach) under varying privacy budgets.

5.2 Data fairness evaluation

To investigate RQ3, Table 3 presents the fairness evaluation of our method in comparison with FairGAN and TabFairGAN, two state-of-the-art generative models explicitly designed to incorporate fairness mechanisms. The results demonstrate that our approach consistently achieves superior fairness outcomes. For instance, in the Pediatric dataset, our method outperformed baseline models by ensuring evenly spread statistical parity across sensitive attributes and achieving the lowest MI values (MI-Gender: 0.0010, MI-Age: 0.0006), the highest Silhouette Score (0.5297), and the lowest DBI score (0.5025) – FairGAN exhibited higher MI values (MI-Gender: 0.1341, MI-Age: 0.1428), lower Silhouette Score (0.0153), and a higher DBI Score (2.0988). Across all datasets, our method consistently attained balanced and evenly spread statistical parity, the lowest DBI scores, and the highest Silhouette Scores in 5 out of 6 datasets. These results highlight the effectiveness of the fairness critic, which dynamically guides the generator to produce bias-free synthetic data by actively penalizing dependencies between sensitive attributes and the generative process.

5.3 Data privacy evaluation

To respond to RQ(5), we report in Table 4 results of privacy, our method consistently yielded low identifiability risk values, ranging from 0.0100 to 0.0221,

Table 3: Comprehensive Analysis of Fairness Metrics Across Baseline Methods.

Dataset	**Method**	**SP(Age)**	**SP(Gender)**	**SP(Ethnicity)**	**MI(Gender)**	**MI(Age)**	**MI(Ethnicity)**	**SS**	**DBI**
HIV	TabFairGAN	-	Balanced	Unbalanced	**0.0000**	-	0.0495	**0.5778**	0.6034
	FairGAN	-	Unbalanced	Unbalanced	0.0517	-	0.0948	0.0926	2.2951
	Our Method	-	Balanced	Balanced	0.0360	-	**0.0656**	0.5527	**0.5349**
Stress	TabFairGAN	Evenly Spread	Balanced	-	0.0102	**0.0000**	-	0.0332	3.8059
	FairGAN	Evenly Spread	Evenly Spread	-	0.0922	0.0000	-	0.1531	2.1520
	Our Method	Balanced	Balanced	-	**0.0009**	0.0040	-	**0.4508**	**1.5409**
Obesity	TabFairGAN	Unbalanced	Balanced	-	**0.0039**	0.0604	-	0.1974	1.5800
	FairGAN	Unbalanced	Unbalanced	-	0.0209	0.0128	-	0.3211	1.7665
	Our Method	Balanced	Balanced	-	0.0346	**0.2027**	-	**0.4166**	**1.1452**
Heart Failure	TabFairGAN	Balanced	Balanced	-	0.0481	0.2340	-	0.3276	0.6575
	FairGAN	Balanced	Unbalanced	-	0.1463	0.4619	-	0.0987	1.4818
	Our Method	Balanced	Balanced	-	**0.0526**	**0.0000**	-	**0.6266**	**0.5150**
Pediatric	TabFairGAN	Evenly Spread	Evenly Spread	-	0.0166	0.0492	-	0.0766	2.0979
	FairGAN	Unbalanced	Balanced	-	0.1341	0.1428	-	0.0153	2.0988
	Our Method	Evenly Spread	Evenly Spread	-	**0.0010**	**0.0006**	-	**0.5297**	**0.5025**
Diabetes	TabFairGAN	Evenly Spread	Balanced	-	0.0601	**0.0000**	-	0.3320	1.2595
	FairGAN	Evenly Spread	Evenly Spread	-	0.1282	0.0000	-	0.0525	2.1114
	Our Method	Evenly Spread	Evenly Spread	-	**0.0005**	0.0028	-	**0.5276**	**0.5003**

indicating strong privacy guarantees across all datasets. Additionally, NNDR scores exceeded 1.0 in five out of six datasets, with higher values (e.g., 3.64 for Obesity) reflecting that synthetic samples were typically more distant from real data points, further supporting privacy preservation.

Table 4: Privacy Risk Metrics Across Datasets

Dataset	**EIR**	**NNDR**
HIV	0.0101	2.0998
Stress	0.0102	1.1012
Obesity	0.0101	3.6430
Heart	0.0100	1.0564
Diabetes	0.0221	1.5492
Pediatric	0.0101	0.0561

5.4 The effect of adversarial debiasing training

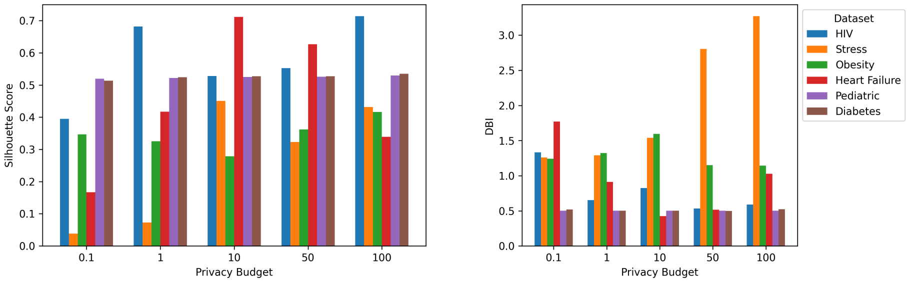
Table 5 presents a comparative analysis of our full architecture, which integrates adversarial debiasing (Adv. Deb.), against a variant that excludes this component (No Adv. Deb.), allowing us to isolate its impact on fairness. The results demonstrate that adversarial debiasing consistently improves fairness metrics, particularly statistical parity and mutual information (MI), as reflected in higher Silhouette Scores and lower DBI values. For instance, in the HIV dataset, our method achieves balanced gender parity and a significantly lower MI of 0.0360, whereas removing adversarial debiasing results in unbalanced parity and a much higher MI of 0.0739. These improvements stem from the dynamic fairness enforcement of the adversarial debiasing critic, which actively penalizes statistical dependencies between sensitive attributes and generated data during training. This ensures that fairness is not a byproduct of data representation but an explicitly optimized objective, effectively preventing demographic biases from being learned and propagated.

Table 5: Ablation Study: Impact of Adversarial Debiasing on Fairness.

Dataset	**Our Method**	**SP(Age)**	**SP(Gender)**	**SP(Ethnicity)**	**MI(Gender)**	**MI(Age)**	**MI(Ethnicity)**	**SS**	**DBI**
HIV	Adv. Deb.	-	**Balanced**	**Balanced**	**0.0360**	-	**0.0453**	**0.5527**	**0.5349**
	No Adv. Deb.	-	Unbalanced	Unbalanced	0.0739	-	0.0587	0.4009	1.2117
Stress	Adv. Deb.	**Balanced**	Balanced	-	**0.0020**	**0.0031**	-	**0.4318**	3.2728
	No Adv. Deb.	Unbalanced	Balanced	-	0.0077	0.0320	-	0.0410	**3.1810**
Obesity	Adv. Deb.	Balanced	Balanced	-	**0.0346**	**0.2027**	-	**0.4166**	**1.1452**
	No Adv. Deb.	Unbalanced	Balanced	-	0.0570	0.1591	-	0.1591	1.7886
Heart Failure	Adv. Deb.	Balanced	Balanced	-	**0.0526**	**0.0000**	-	**0.6266**	**0.5150**
	No Adv. Deb.	Balanced	Balanced	-	0.0900	0.3022	-	0.3022	1.0797
Pediatric	Adv. Deb.	Evenly Spread	Evenly Spread	-	**0.0010**	**0.0006**	-	**0.5297**	**0.5025**
	No Adv. Deb.	Evenly Spread	Evenly Spread	-	0.0186	0.4161	-	0.4161	0.6323
Diabetes	Adv. Deb.	Evenly Spread	Evenly Spread	-	**0.0008**	0.3185	-	0.3185	**0.5000**
	No Adv. Deb.	Evenly Spread	Evenly Spread	-	0.0839	0.3185	-	0.3185	0.7777

5.5 The effect of privacy on fairness

To answer RQ4, we assess the tradeoff between fairness and privacy. We analyzed the evolution of Silhouette Scores and Davies-Bouldin Index values under varying privacy budgets and present them in Figure 5. Our results indicate that as the privacy budget increases, Silhouette Scores improve, reflecting enhanced cohesion, while DBI values decrease, signifying enhanced fairness. For instance, in the HIV dataset, the Silhouette Score rises from 0.3953 at $\varepsilon = 0.1$ to 0.7143 at $\varepsilon = 100$, while DBI decreases from 1.3341 to 0.5900, demonstrating improved fairness level with relaxed privacy constraints. However, certain anomalies suggest that excessive noise can introduce instability in clustering quality. Notably, in the Stress dataset, DBI spikes from 1.5409 at $\varepsilon = 10$ to 3.2728 at $\varepsilon = 100$, while the Silhouette Score drops from 0.4508 to 0.4318. Despite leveraging zCDP for tighter privacy bounds via Rényi divergence, these results highlight the nuanced sensitivity of tuning the zCDP privacy budget. While zCDP reduces instability compared to traditional DP, the relationship between privacy strength (ρ) and fairness remains irregular, highlighting the delicacy of calibrating zCDP. This behavior is partly influenced by the fairness critic, which guides the model to generate balanced representations across subgroups. As zCDP introduces noise, the fairness critic retains more signal to enforce equitable data generation. However, when the privacy budget is too tight (low ρ), the added noise limits the critic's ability to discern and correct subgroup disparities, weakening fairness enforcement. Conversely, with a looser budget, the critic can more effectively align distributions across sensitive attributes, leading to lower DBI scores.



(a) Silhouette Scores across different privacy budgets. Higher values indicate better-defined clusters, showing the trade-off between privacy and clustering quality.

(b) Davies-Bouldin Index (DBI) under different privacy budgets. Lower values indicate better clustering structure and separation, highlighting the fairness-privacy trade-off.

Fig. 5: Comparison of clustering performance under different privacy budgets using Silhouette Score and Davies-Bouldin Index (DBI). The results illustrate the impact of privacy constraints on clustering structure and fairness.

5.6 The effect of clustering on data quality

Figure 6 shows that clustering sensitivity to the number of clusters (k) depends on dataset characteristics. HIV, with many categorical features, improves steadily with k, peaking at 25 clusters, suggesting complex subgroup structures. Heart, being small and mostly numerical, is highly sensitive—peaking at $k = 20$ then dropping—indicating the risk of excessive cluster formation, leading to the loss of meaningful patterns. Stress dataset starts with poor clustering at $k = 10$ (0.04), but improves significantly by $k = 20$, suggesting that more clusters are needed to meaningfully separate complex patterns in the data. The drop in performance at $k = 23$ and 25 indicates over-segmentation, where noise may be mistaken for distinct groups. In contrast, Pediatric and Diabetes, both high-dimensional and largely numerical, show stable scores across k, indicating consistent group patterns. Obesity shows moderate improvement. Overall, smaller or more categorical datasets are more sensitive to k, while large numerical datasets are more robust.

6 Conclusion

We proposed a novel hybrid generative framework that enforces fairness and privacy in unsupervised synthetic data generation. By integrating a clustering-based Variational Autoencoder with a Wasserstein GAN with Gradient Penalty, our approach structures latent representations while ensuring privacy through zero-Concentrated Differential Privacy. Adversarial debiasing via the Fairness Critic mitigates bias without requiring explicit labels. Extensive experiments across

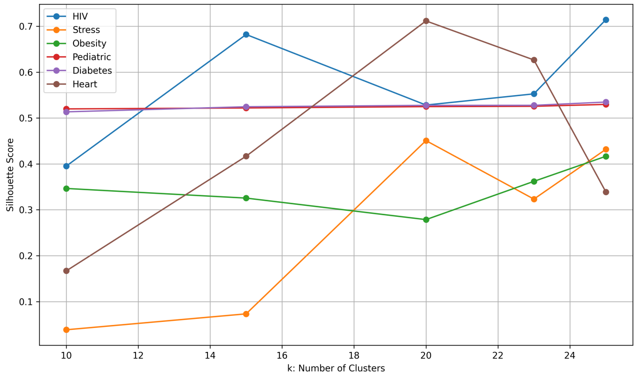


Fig. 6: Silhouette Score Across Six Datasets.

multiple healthcare datasets demonstrate our method's superiority over existing fairness-aware and privacy-preserving generative models. By leveraging various data realism and fairness metrics, we provided a rigorous and interpretable evaluation of generative quality, bias mitigation, and privacy preservation. Our results highlight the effectiveness of our approach in generating high-quality, fair, and privacy-compliant synthetic data across varying privacy budgets in unsupervised settings. As future work, we aim to explore adaptive clustering techniques to dynamically adjust cluster formation and contrastive learning to further enhance bias mitigation.

7 Acknowledgments

This work was supported by the France 2030 grants RHU RECORDS (ANR-18-RHUS-0004) and IHU PROMETHEUS (ANR-23-IAHU-0004) and the iRECORDS project, funded by ERA PerMed (JTC_2021).

References

1. Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci*, 6(1), 3. doi:10.3390/sci6010003
2. Rajabi, A., & Garibay, O. O. (2021). TabFairGAN: Fair Tabular Data Generation with Generative Adversarial Networks. *arXiv preprint arXiv:2109.00666*. arXiv:2109.00666
3. Teo, C., Abdollahzadeh, M., & Cheung, N. M. M. (2023). On measuring fairness in generative models. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 10644–10656).
4. Xie, L., Lin, K., Wang, S., Wang, F., & Zhou, J. (2018). Differentially Private Generative Adversarial Network. *arXiv preprint arXiv:1802.06739*. arXiv:1802.06739
5. Torfi, A., Fox, E. A., & Reddy, C. K. (2022). Differentially private synthetic medical data generation using convolutional GANs. *Information Sciences*, 586, 485–500.

6. Movva, R., Koh, P. W., & Pierson, E. (2024). Using unlabeled data to enhance fairness of medical AI. *Nature Medicine*, 30, 944–945. doi:10.1038/s41591-024-02892-0
7. Xu, D., Yuan, S., Zhang, L., & Wu, X. (2018). FairGAN: Fairness-aware Generative Adversarial Networks. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 570–575). IEEE.
8. Xu, D., Yuan, S., Zhang, L., & Wu, X. (2019). FairGAN+: Achieving fair data generation and classification through generative adversarial nets. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 1401–1406). IEEE.
9. Bun, M., & Steinke, T. (2016). Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. *arXiv preprint arXiv:1605.02065*. doi:10.1145/3588433
10. Deshmukh, G., & Naik, A. (2023). Biases and Fairness in Deep Learning Models: A Survey on Inculcating Fairness in Generative Models. In *2023 7th International Conference On Computing, Communication, Control And Automation (ICCUBEA)* (pp. 1–39). doi:10.1109/ICCUBEA58933.2023.10391962
11. Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care? *AMA Journal of Ethics*, 21(2), 167–179.
12. Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*. arXiv:1808.00023
13. Zamzam, O., Akrami, H., & Leahy, R. M. (2022). Learning From Positive and Unlabeled Data Using Observer-GAN. *arXiv preprint arXiv:2208.12477v2*. arXiv:2208.12477
14. Zhao, Y., Xu, Q., Wen, P., Jiang, Y., & Huang, Q. (2022). Dist-PU: Positive-Unlabeled Learning from a Label Distribution Perspective. *arXiv preprint arXiv:2212.02801*. arXiv:2212.02801
15. De Paolis Kaluza, M. C., Jain, S., & Radivojac, P. (2023). An Approach to Identifying and Quantifying Bias in Biomedical Data. *Pacific Symposium on Biocomputing*, 28, 311–322.
16. Ghosheh, G. O., Li, J., & Zhu, T. (2023). A Survey of Generative Adversarial Networks for Synthesizing Structured Electronic Health Records. *ACM Computing Surveys*, 55(8). doi:10.1145/3636424
17. Ros, F., Riad, R., & Guillaume, S. (2023). PDBI: A Partitioning Davies-Bouldin Index for Clustering Evaluation. *Neurocomputing*, 528, 178–199. doi:10.1016/j.neucom.2023.01.043
18. Kang, J., Xie, T., Wu, X., Maciejewski, R., & Tong, H. (2022). InfoFair: Information-Theoretic Intersectional Fairness. In *2022 IEEE International Conference on Big Data (Big Data)* (pp. 1455–1464). doi:10.1109/BigData55660.2022.10020588
19. Jordon, J., Yoon, J., & Van Der Schaar, M. (2018). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*. openreview.net
20. Wang, H., Bi, J., Hua, M., Yan, K., & Afshari, A. (2025). Semi-supervised CWGAN-GP modeling for AHU AFDD with high-quality synthetic data filtering mechanism. *Building and Environment*. doi:10.1016/j.buildenv.2024.112265
21. Ma, C., Zhao, R., Xiao, X., Xie, H., Wang, T., Wang, X., Zhang, H., & Shen, Y. (2025). CAD-VAE: Leveraging Correlation-Aware Latents for Comprehensive Fair Disentanglement. *arXiv preprint arXiv:2503.07938*. arXiv:2503.07938