Int J Comput Intell Syst

https://doi.org/10.1007/s44196-026-01173-7

Article in Press

Probabilistic Versus Deep Generative Models: A Fairness Centred Evaluation of Synthetic Healthcare Tabular Data

Received: 8 September 2025

Accepted: 15 January 2026

Published online: 26 February 2026

Cite this article as: Alattal D., Draghi B., Myles P. *et al.* Probabilistic Versus Deep Generative Models: A Fairness Centred Evaluation of Synthetic Healthcare Tabular Data. *Int J Comput Intell Syst* (2026). https://doi.org/10.1007/s44196-026-01173-7

Dima Alattal, Barbara Draghi, Puja Myles, Richard Branson & Allan Tucker

Probabilistic Versus Deep Generative Models: A Fairness Centred Evaluation of Synthetic Healthcare Tabular Data

Probabilistic vs Deep Generative Models: A Fairness Centred Evaluation of Synthetic Healthcare Tabular Data

Dima Alattal[1], Barbara Draghi[2], Puja Myles, Richard Branson,[3,4†], Allan Tucker[5*]

[1,2,5*]Computer Science Department, Brunel University London, London, UK.
[3,4] Medicine and Healthcare products Regulatory Agency, London, UK.

*Corresponding author(s). E-mail(s): allan.tucker@brunel.ac.uk;
Contributing authors: dima.alattal2@brunel.ac.uk;
barbara.draghi@brunel.ac.uk; puja.myles@mhra.gov.uk,
richard.branson@mhra.gov.uk;
†These authors contributed equally to this work.

Abstract

Synthetic data offers a promising avenue for addressing privacy, scarcity, and fairness challenges in healthcare datasets. However, there is limited evaluation of how different generation methods balance fidelity, utility, and fairness, particularly for underrepresented subgroups. This study addresses this gap by comparing representative generative modelling techniques, both probabilistic and deep approaches, that are popular in the research literature. We empirically evaluate BayesBoost, CTGAN, TVAE, CopulaGAN, and DECAF on two healthcare datasets containing numerical, binary, and categorical features. Each model's performance is assessed along three axes: data fidelity, machine learning utility, and fairness, using Accuracy Parity, Equalised Odds, and Predictive Rate Parity. Results show that BayesBoost consistently achieved superior fidelity, utility, and fairness preservation, particularly when paired with Random Forest classifiers, achieving around *60–63% higher* downstream utility than GAN-based deep generative baselines (e.g., Random Forest accuracy up to *0.88* with BayesBoost versus *0.54–0.55* for GAN-based methods). Deep generative models, while

effective in capturing complex structures, often degraded fairness, especially for underrepresented groups, with equalised odds deviating by over *100%* from the ideal parity value of *1.0* in some settings. The Variational Autoencoder outperformed other deep generative models in fairness preservation, especially for equalised odds, although with some reduction in fidelity and utility. Overall, these findings suggest that synthetic data generation for healthcare must move beyond fidelity evaluations to explicitly assess fairness and subgroup impacts, with probabilistic models such as BayesBoost showing strong potential for ethical deployment, while deep generative models require further adaptation for fairness-sensitive applications.

Keywords: Synthetic Data Generation, Tabular Data, Fairness in Machine Learning, Healthcare Data, Generative Models, Data Fidelity, Bias Mitigation, BayesBoost, GAN, VAE

1 Introduction

Synthetic tabular data generation has emerged as a promising solution to critical challenges in healthcare, including data privacy, scarcity, and bias mitigation [1]. The growing digitalisation of healthcare and the adoption of electronic health records have created new opportunities for data-driven research but have also intensified concerns around privacy, security, and data-sharing.

Traditional privacy-preserving approaches, such as de-identification and anonymisation, aim to protect sensitive information by removing direct identifiers and introducing random noise [2, 3]. However, these techniques can degrade data fidelity, compromise analytical validity, and risk introducing new biases into downstream analyses.

An alternative that can potentially overcome these limitations involves the generation of fully synthetic data as a substitute for real datasets. Synthetic data generation produces artificial records that reproduce the statistical structure of the original data while avoiding the disclosure of identifiable information. In healthcare, where most information is stored in structured tabular form describing patients through mixed numerical, categorical, and binary variables, generating realistic synthetic

data requires capturing complex conditional dependencies and maintaining clinical plausibility.

Previous studies have explored the generation of synthetic patient-level datasets from aggregated population statistics [4, 5], balancing confidentiality and statistical representativeness. Probabilistic models, particularly Bayesian Networks (BNs), have demonstrated success in replicating complex clinical associations and joint distributions [6, 7]. Building on this foundation, BayesBoost integrates boosting techniques to better address representation bias and enhance synthetic data quality for underrepresented groups [8].

Recent advances in machine learning have expanded synthetic data generation capabilities further. Deep generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) [9, 10] offer powerful tools for capturing non-linear feature interactions without requiring explicit distributional assumptions. Models like CTGAN, TVAE and CopulaGAN have been adapted to tabular data, demonstrating flexibility in complex data settings [11]. Nevertheless, challenges persist. Tabular healthcare datasets often contain a mix of continuous, binary, and categorical features with high sparsity, imbalance, and complex dependencies, which generative models struggle to replicate faithfully.

Fairness in synthetic data generation is a significant concern, particularly within healthcare, where biased data can result in unequal outcomes for various patient groups. Ensuring fairness involves not only achieving equitable representation of subgroups within the synthetic data but also guaranteeing fair and comparable outcomes for underrepresented groups in machine learning models trained on this data. However, this is a challenging task, as generative models may unintentionally reproduce, amplify biases present in the original datasets or even introduce new biases. To address these issues, robust evaluation frameworks are essential, emphasising both fairness and ethical considerations.

This study addresses these gaps through an empirical evaluation of five synthetic data generation methods—BayesBoost, CTGAN, TVAE, CopulaGAN, and DECAF—on healthcare datasets comprising numerical, binary, and categorical features for binary classification tasks. We systematically assess how these methods preserve statistical fidelity, predictive utility, and fairness, with a particular focus on underrepresented groups.

While previous surveys have evaluated synthetic data generation techniques with respect to fidelity, utility, and privacy [9, 12–16], few have explicitly addressed fairness preservation, especially in healthcare. Moreover, no previous studies have compared deep generative models against BayesBoost. a probabilistic approach specifically designed to mitigate bias in tabular data generation. While DECAF and BayesBoost incorporate bias mitigation mechanisms, CTGAN, TVAE and CopulaGAN are primarily designed to capture complex tabular structures without explicit fairness objectives. By evaluating fidelity, machine learning utility, and fairness in tandem, our study provides a comprehensive, comparative assessment of the capabilities and limitations of these approaches. The findings emphasise that fairness-conscious design, structured evaluation frameworks, and transparent reporting must be prioritised to ensure the ethical deployment of synthetic data systems in healthcare applications.

In summary, this paper makes three main contributions:

1. **Comprehensive evaluation framework:** We introduce an integrated assessment framework to jointly evaluate *fidelity*, *predictive utility*, and *fairness* in synthetic tabular data, enabling systematic comparison across modelling paradigms.

2. **Empirical comparison of probabilistic and deep generative models:** We conduct a rigorous experimental analysis of five representative approaches (BayesBoost, CTGAN, TVAE, CopulaGAN, and DECAF) using healthcare datasets with mixed numerical, categorical, and binary features.

3. **Insights on fairness preservation and ethical deployment:** We provide new evidence on how probabilistic and deep models differ in their ability to preserve subgroup fairness, highlighting the strengths and limitations of the evaluated approaches for fairness-sensitive healthcare applications.

1.1 Related Work

Synthetic data holds considerable value in healthcare, offering a range of significant benefits that address both practical and ethical challenges. By augmenting existing datasets, synthetic data enhances the performance of machine learning models through increased data volume and diversity, which is particularly beneficial when dealing with rare diseases or underrepresented patient populations. It also improves data accessibility, enabling researchers to overcome limitations related to data sharing restrictions. Beyond these advantages, synthetic data provides a powerful privacy-preserving solution by mimicking the statistical patterns of real data while excluding any personally identifiable information [17].

In addition, synthetic health records play a pivotal role in in-silico clinical trials, which use patient-specific models to generate virtual cohorts for evaluating the safety and effectiveness of new drugs and medical devices [18]. By using synthetic data in such trials, researchers can also refine trial designs and make predictions at both population and individual levels, thus improving the chances of success [19].

1.1.1 Probabilistic Generative Models

BayesBoost, a type of probabilistic data generation method, addresses biases in the original dataset by targeting difficult-to-predict or under-represented samples. This is particularly important as biases in real data often propagate into synthetic data, potentially compromising its utility and fairness. BayesBoost identifies these under-represented instances and employs a boosting approach to mitigate the issue. Boosting, a machine learning technique that combines multiple weak models to produce a

stronger one, is used here to over-sample the under-represented cases. This process ensures a more balanced and representative distribution of subgroups in the synthetic dataset. The synthetic data generation step in BayesBoost relies on Bayesian networks to model the relationships and distributions in the original data. Bayesian networks are probabilistic graphical models that use directed acyclic graphs to represent variables and their conditional dependencies [8].

Bayesian networks are widely recognised for their capability to model probabilistic relationships and perform inference tasks, allowing the representation of causal dependencies within data. These characteristics make them particularly valuable for understanding variable interactions and incorporating uncertainty into predictive models. Additionally, their flexibility in handling non-linear data patterns enhances their applicability in domains where decision-making and risk assessment are critical. However, the application of Bayesian networks to large-scale, high-dimensional datasets and complex distributions presents significant challenges. Learning intricate dependencies in such contexts is computationally demanding and often requires complex structuring methods [13].

1.1.2 Deep Learning Generative Models

Deep learning-based data generation methods, particularly GANs, have undergone significant evolution to address the unique challenges posed by various data types, including tabular data.Vanilla GANs, the foundational GAN architecture, consist of a generator and discriminator that engage in an adversarial framework to create synthetic data. While effective for image data, Vanilla GANs face limitations when applied to tabular data due to the complexity of modelling mixed data types (categorical and numerical) and the intricate dependencies inherent in such datasets [20].

To overcome these challenges, specialized GAN variants for tabular data have been introduced. One notable example is Conditional GANs (CTGANs), which incorporate specific techniques like mode-specific normalization to better handle the characteristics

of tabular datasets. CTGANs enable conditional data generation, allowing synthetic data to be generated based on specific labels or features. This feature is particularly advantageous in structured data settings, where controlling the generation process based on attributes such as disease type or patient demographics enhances the realism and practical utility of the synthetic data [21].

Copula GANs represent another advancement in tabular data synthesis, using copula theory to model interdependencies among variables. This approach allows for more accurate capture of relationships in high-dimensional data, particularly when dealing with mixed data types (continuous and categorical), thereby offering improvements over Vanilla GANs [21].

DECAF takes a different approach by prioritizing fairness and bias mitigation in data generation. DECAF explicitly incorporates the data-generating process into its framework as a structural causal model embedded within the generator's input layers. By reconstructing each variable based on its causal parents, DECAF ensures that the generated data faithfully reflects the underlying causal relationships in the original dataset. This feature enables DECAF to produce synthetic data that minimizes biases while preserving critical causal structures [12].

VAEs, another variation of deep generative models, are based on the autoencoder architecture, comprising an encoder and a decoder. The encoder transforms input data into a compressed, continuous latent space represented by probabilistic latent variables. The decoder then reconstructs the data from this latent space, enabling VAEs to generate samples that align with the original data distribution. Unlike traditional autoencoders, VAEs utilise probabilistic encoding, which maps input data into distributions rather than fixed points in the latent space. This probabilistic framework facilitates smooth transitions between generated samples, making VAEs particularly effective in producing diverse synthetic datasets that match the distribution of the original data.

However, VAEs, like other deep generative models such as GANs, face several limitations. They can be computationally intensive, particularly when applied to large and complex datasets. One significant challenge is model collapse, where the generator fails to capture the full variability of the data, resulting in limited sample diversity and repetitive outputs. Such issues undermine the model's ability to represent the complexity of real-world data effectively. Moreover, the quality and fidelity of synthetic data generated by VAEs depend heavily on the adequacy and balance of the training data. If the training dataset is insufficient, biased, or non-representative, the synthetic data may deviate from the true underlying distribution of the original dataset. This discrepancy compromises both the fidelity and the utility of the synthetic data, particularly in sensitive applications.

1.1.3 Fairness and Bias Mitigation in Synthetic Data

While generative models hold significant promise for advancing the healthcare system, they also raise ethical concerns regarding their generalisability across diverse populations. These concerns are magnified when data is skewed or underrepresented, leading to harmful biases related to age, race, or gender. In such cases, synthetic data generation can serve as a potential mitigation strategy by addressing these gaps and ensuring better representation within datasets [17].

Fairness in AI systems is a major concern, as an AI system may exhibit unfair behaviour when it extends or withholds opportunities, resources, or information unequally across different groups. Furthermore, disparities in quality of service—where the system performs well for one group but inadequately for another—can exacerbate existing inequities. In healthcare, such biases can significantly impact treatment outcomes, leading to unequal access and care for specific racial, gender, or demographic groups. These disparities are not just inequitable but can result in life-threatening consequences.

Biases in AI systems often stem from multiple sources, including the use of imbalanced or underrepresented data to train generative models, inherent algorithmic biases, and the ways in which generative and predictive AI models are deployed in practice [22, 23]. Biases in the original data are frequently carried over to synthetic datasets generated by AI models, perpetuating the inequities in subsequent analyses. This can lead to predictive machine learning models trained on synthetic data that exhibit a lack of fairness, ultimately causing harm to individuals. For instance, biased models may result in misdiagnoses, inappropriate treatment plans, or unequal access to vital medical resources. Addressing these challenges requires targeted efforts to mitigate biases at every stage of the AI pipeline from data collection and generation to model training and deployment to ensure equitable and ethical healthcare practices [23].

Addressing biases in synthetic data generation can involve multiple strategies, including eliminating biases in the original datasets, balancing data during the generation process, or adapting generative models to account for existing biases. One such approach, BayesBoost, is designed to tackle biases from a representation perspective by applying boosting techniques to underrepresented sub-populations identified through model performance metrics [8]. By focusing on these sub-populations, BayesBoost aims to create a more balanced synthetic dataset. However, its effectiveness and consistency in producing unbiased and informative results require further empirical validation. Decaf, approaches the issue differently by incorporating causal frameworks. Instead of balancing the initial dataset, Decaf employs causal graphs to disconnect the direct relationship between sensitive attributes (e.g., gender, race) and the target variable while preserving the sensitive attributes in the data. This ensures that the generated synthetic data does not encode direct causal relationships between sensitive attributes and the target variable, thereby preventing sensitive attributes from influencing classification outcomes. This approach allows for maintaining fairness without

the need to exclude sensitive variables entirely, offering a nuanced way of addressing biases in synthetic data generation [12].

Fairness in machine learning has been a key focus of numerous studies, with particular attention given to developing measures that identify biases within datasets and model outcomes. These measures assess fairness either across demographic groups or at an individual level. For group-based fairness, prominent metrics include Demographic Parity (DP) [24], Predictive Rate Parity (PRP) [25], Equalized Odds (EO), Equal Opportunity [26], and Accuracy Equality [27]. These measures evaluate whether different demographic groups, such as those defined by race, gender, or age, have an equal opportunities of receiving favorable outcomes—for instance, being correctly diagnosed in healthcare scenarios.

On the individual level, fairness measures assess whether a model provides consistent outputs for individuals differing only in sensitive attributes, such as gender or ethnicity [24]. This perspective considers an algorithm to be fair if its predictions are independent of such attributes when all other factors are identical. More extensive fairness definitions and measures have been formalized and reviewed extensively in [27].

2 Experiments & Methods

This section outlines the experimental framework established to evaluate five state-of-the-art synthetic tabular data generation methods: the deep generative models CopulaGAN, CTGAN, DECAF, and VAE, as well as the probabilistic model BayesBoost. The primary focus is to assess these models' performance in terms of fidelity, utility, and fairness, particularly in healthcare datasets characterised by mixed data types (numerical, binary, and categorical) and underrepresented sensitive attributes.

We detail the train-test splitting strategy, hardware and software configurations, data preprocessing steps, and model-specific configurations adopted in the experiments.

2.1 Datasets

Our experiments employed two tabular datasets, each designed for binary classification tasks within the healthcare domain. The first dataset, derived from synthetic CPRD primary care data [28], is a high-fidelity synthetic dataset focusing on cardiovascular disease risk factors. It includes variables such as smoking behaviour, age, and chronic conditions associated with cardiovascular health. The version used comprises 10,000 synthetic individuals, randomly sampled from the synthetic CPRD data, with a binary target variable indicating the occurrence or absence of a heart attack.

The second dataset is a publicly available resource from Kaggle, focusing on diabetes diagnosis as a binary target variable. It comprises detailed health records for 1,879 patients, including demographic information, lifestyle factors, and medical history [29]. A more detailed description of both datasets is provided in Table 1.

Both datasets were selected specifically for their relevance to subgroup fairness analysis, given their inherent representation biases in sensitive attributes and the variation in subgroup-specific fairness metrics. Notably, they range from datasets exhibiting negligible fairness concerns to those presenting clearer disparities. While maintaining reasonable accuracy parity across demographic subgroups, both datasets show moderate or minimal fairness issues regarding Predictive Rate Parity and Equalised Odds, as detailed in Tables 4 and 5. These characteristics make them particularly suitable for evaluating the ability of synthetic data generation models to preserve, degrade, or enhance fairness relative to the real data.

Table 1: Dataset descriptions used in the experiments

Data	CVD	Diabetes
Number of instances	10,000	1,879
Number of features	22	44
Numerical features	5	21
Binary features	13	19
Categorical features	4	4
Sensitive attribute	Ethnicity (6 groups)[1]	Ethnicity (4 groups)[2]
Protected group	Group 0	Group 2
Target variable	Heart attack	Diabetes diagnosis

[1]CVD dataset sensitive attribute groups: 0: Asian; 1: Black; 2: Mixed; 3: Other; 4: Unknown; 5: White.

[2]Diabetes dataset sensitive attribute groups: 0: Caucasian; 1: African American; 2: Asian; 3: Other.

2.2 Experimental Settings

Among the most widely used deep learning generative models for tabular data, we evaluated several state-of-the-art approaches extensively explored in recent studies. These models include CopulaGAN, CTGAN, TVAE, and DECAF. In addition, we compared these deep generative models against BayesBoost, a probabilistic approach proposed as a potential solution to address bias issues in synthetic tabular data generation.

To mitigate the risk of information leakage, the ground truth datasets were partitioned into an 80/20 train-test split. The synthetic data generation process was performed exclusively on the training set, ensuring complete independence of the test set for subsequent evaluation. Test data were randomly sampled to preserve the distributions of both the sensitive attribute and the target variable, ensuring adequate subgroup representation for fairness assessment.

Each generative model produced synthetic datasets across 150 independent iterations, each utilising a distinct random seed. Performance metrics were averaged across these iterations to account for variability and to ensure robust and reliable evaluation results.

Various libraries and model-specific implementations were employed to train and evaluate the models. BayesBoost was developed in RStudio using the `bnlearn` library,

while the Synthetic Data Vault (SDV) library was used to implement CTGAN, CopulaGAN, and TVAE models. The original implementation was used for the DECAF model. To ensure consistency across experiments, default settings were adopted for batch size and number of training epochs across all deep learning models. Given the computational demands and complexity associated with hyperparameter tuning in deep generative models, all hyperparameters were kept at their default values, recognising that this choice may not exhaustively explore the parameter search space.

All datasets underwent comprehensive preprocessing to ensure consistency and comparability prior to training the generative models. For BayesBoost, which requires categorical data, all features were discretised before model training. To maintain uniformity in model comparisons, a standardised preprocessing approach was applied across all deep generative models. Categorical features were initially converted into numerical values using the LabelEncoder from the scikit-learn library. Following this, both categorical and numerical features were scaled to the range [0, 1] using the MinMaxScaler, also from scikit-learn. The MinMaxScaler was selected for data transformation after preliminary trials demonstrated the best performance compared to alternative scaling methods.

To prevent data leakage, the scaler was fitted exclusively on the training data and then applied to both training and test sets. This scaling procedure was consistently repeated for each model across all 150 iterations, with a new scaler implemented in each iteration. The generated data was subsequently reverse-transformed to its original format to enable accurate data fidelity comparisons.

For the DECAF model, we generated the direct causal graph using the PC algorithm in GEnie Modeler and applied the no-debias (Decaf ND) approach [30]. This method infers the causal relationships within GANs without removing the causal edges between the target variable and sensitive attributes. We specifically chose not to remove any causal relationships between ethnicity and health outcomes, hence not

employing other DECAF variations such as DECAF FTU, DECAF DP, or DECAF CU. This choice was based on the critical role that ethnicity plays in healthcare diagnosis; by preserving the influence of ethnicity on medical outcomes, we aimed to support fair and accurate predictions across all ethnic groups, rather than eliminating these relationships.

For each synthetic dataset generated, metrics for fidelity, utility, and fairness were calculated to comprehensively evaluate the performance of each generative model. After computing these metrics for each individual dataset, the average scores across the 150 iterations were used to assess each model's overall performance in terms of the selected evaluation metrics, yielding a balanced and robust assessment of each model's capabilities.

All experiments were performed on a system equipped with an Intel Evo CPU @ 3.80 GHz, 32 GB RAM, and an NVIDIA GeForce RTX 4060 GPU with 16 GB of memory. Python (version 3.12) was used for the majority of models, while R Studio was utilised for the BayesBoost experiments.

2.3 Evaluation Metrics

2.3.1 Data Fidelity

Data fidelity reflects the extent to which synthetic data accurately replicates the statistical characteristics and patterns observed in the original dataset. To evaluate fidelity, we conducted both bivariate and multivariate analyses, assessing categorical and numerical features using a range of complementary metrics.

For bivariate analysis, separate metrics were applied to categorical and numerical variables. In the case of categorical variables, *Category Coverage* assessed whether the synthetic data captured all unique categories present in the real data. High category coverage indicates that the synthetic data retains the diversity of categorical

values, ensuring representativeness across all subgroups. Additionally, the *TV Complement*, based on the total variation distance, quantified the similarity between the distributions of categorical values in the real and synthetic datasets. A higher TV complement score suggests closer alignment between the two distributions, indicating superior fidelity.

For numerical variables, three metrics were employed. *Boundary Adherence* evaluated whether the synthetic data respected the minimum and maximum values observed in each real data column, ensuring the absence of unrealistic or out-of-bound values. *Range Coverage* assessed whether the synthetic data spanned the full range of real data values, thus preserving the variability and representativeness of numerical features. Finally, the *KS Complement*, derived from the Kolmogorov-Smirnov statistic, measured the similarity between the marginal distributions of numerical variables in the real and synthetic datasets. Higher KS complement values indicate better alignment and therefore higher fidelity.

To ensure robustness, all fidelity scores were averaged across the 150 synthetic datasets generated for each model. This averaging mitigated variability introduced by random seed selection and provided a more reliable measure of model performance.

For multivariate fidelity assessment, we evaluated the ability of synthetic datasets to preserve inter-variable relationships by computing the correlation matrices of numerical features for both real and synthetic datasets. Each synthetic dataset, generated across 150 iterations per model, was compared with the real data correlation matrix using the Frobenius norm distance. This metric quantifies the dissimilarity between two correlation matrices, where lower Frobenius distances indicate greater similarity to the real data structure, and higher values suggest greater divergence. To enhance reliability, Forbes distances were averaged across all iterations for each model, providing a comprehensive measure of each model's ability to preserve multivariate dependencies.

2.3.2 ML Utility

To evaluate the utility of synthetic data, we assessed its predictive utility—namely, the ability of machine learning models trained on synthetic data to replicate the predictive performance of models trained on real data. Predictive utility was evaluated by training three machine learning models—Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB)—on synthetic datasets and testing them on the real test set. The test set, comprising 20% of the original data, was held out from all synthetic data generation processes to ensure independent and unbiased evaluation.

The classification task involved binary prediction of health conditions. Models were trained using 5-fold cross-validation to mitigate sampling variance and ensure robust performance estimation. Model performance was evaluated using recall, precision, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC).

These metrics were selected due to their clinical relevance: recall reflects the ability to correctly identify true positive cases, which is crucial in healthcare diagnostics to avoid missed cases; precision measures the accuracy of positive predictions, helping to minimise false positives and prevent unnecessary interventions; F1-score balances precision and recall, providing an overall assessment where both types of errors are consequential; and AUC-ROC evaluates the model's overall discriminative ability, offering insight into predictive reliability across various decision thresholds.

2.3.3 Fairness

To assess fairness, we employed three widely used group fairness metrics: Accuracy Parity, Equalised Odds, and Predictive Rate Parity. These metrics were applied to evaluate and compare the fairness of machine learning models across subgroups within the sensitive attribute. The analysis specifically focused on the most underrepresented subgroup—referred to as the "protected subgroup"—in each dataset. In the CVD

dataset, the protected subgroup was Group 0, while in the Diabetes dataset, it was Group 2, both corresponding to individuals of Asian ethnicity.

Accuracy Parity is achieved when the prediction accuracy for the protected subgroup is equivalent to that of the unprotected subgroups. This metric measures the probability of making correct predictions (whether positive or negative) across different groups. In this context, it ensures that individuals with or without a health condition are classified with comparable overall accuracy, regardless of subgroup membership. Formally, this can be expressed as: $P(d = Y, SA = PS) = P(d = Y, SA = SG)$ where d represents the model's prediction, Y is the true label, SA is the sensitive attribute, PS is the protected subgroup, and SG is any other subgroup. In our evaluation, Accuracy Parity was considered satisfied when the classification accuracy was similar between the protected and unprotected subgroups.

Predictive Rate Parity assesses whether precision (or positive predictive value, PPV) is consistent across subgroups. PPV indicates the likelihood that a positive prediction corresponds to an actual positive case. Fairness under this metric requires that the proportion of true positives among all predicted positives is equivalent for the protected and unprotected subgroups $P(Y = 1 \mid d = 1, SA = PS) = P(Y = 1 \mid d = 1, SA = SG)$. This metric also indirectly ensures equality in false discovery rates (FDR), which measure the proportion of false positives among all predicted positives: $P(Y = 0 \mid d = 1, SA = PS) = P(Y = 0 \mid d = 1, SA = SG)$. In our experiments, this implies that for both protected and unprotected subgroups, the probability of a patient being accurately classified as having a health condition should be the same. This metric ensures that individuals receiving positive predictions from different subgroups have equal confidence in their prediction accuracy.

Equalised Odds ensures that a classifier's true positive rate (TPR) and false positive rate (FPR) are approximately equal across subgroups. This metric evaluates the consistency of correct and incorrect predictions between the protected and unprotected

subgroups. Specifically, patients with and without a health condition should have equal probabilities of being correctly or incorrectly classified, irrespective of their sensitive attribute values. Mathematically, for a binary outcome $P(d = 1 \mid Y = i, SA = PS) = P(d = 1 \mid Y = i, SA = SG)$. Where, i denotes the binary health condition (present or absent), and d=1 indicates a positive prediction .

In our experiments, a model was deemed to maintain fairness if the values of the fairness metrics (Accuracy Parity, Equalised Odds, and Predictive Rate Parity) for the protected subgroup were approximately equal to those of other subgroups, ideally close to a ratio of 1. Significant deviations from this benchmark highlighted fairness concerns. A ratio greater than 1 indicated that the unprotected subgroup benefited more from model performance relative to the protected group, while a ratio below 1 suggested that the protected subgroup had relatively better outcomes compared to the other subgroup.

A complete summary of the experimental framework, including datasets, preprocessing, generative model configurations, evaluation metrics, and computational environment, is provided in Table 2. This table consolidates all methodological components and serves as a reference for the subsequent results and discussion.

3 Results

3.1 Data Fidelity

Figure 1 presents the statistical fidelity results for the CVD and Diabetes datasets across five synthetic data generation models: BayesBoost, CTGAN, VAE, DECAF, and CopulaGAN. The evaluation includes five complementary metrics: Category Coverage, TV Complement, Range Coverage, Boundary Adherence, and KS Complement. These metrics assess how closely the synthetic data replicate key statistical properties of the real datasets. Category Coverage and Range Coverage measure the extent to which categorical and numerical feature values in the synthetic data overlap with

Table 2: Summary of experimental design, models, and evaluation settings

Component	Description	Details / Parameters	Purpose / Output
Datasets	Two binary classification healthcare datasets	CVD (10,000 samples, 22 features); Diabetes (1,879 samples, 44 features)	Assess model generalisability across domains
Sensitive attribute	Ethnicity (categorical, multiple subgroups)	CVD: 6 groups (protected = Group 0, Asian); Diabetes: 4 groups (protected = Group 2, Asian)	Enable subgroup fairness evaluation
Train–test split	Stratified 80/20 split	Preserves distributions of sensitive and target variables	Prevents data leakage and ensures independent evaluation
Preprocessing	Encoding and scaling	LabelEncoder + MinMaxScaler [0,1]; scaler fitted on train set only	Standardised input for all models
Models	Five state-of-the-art methods	CopulaGAN, CTGAN, TVAE, DECAF (Python/SDV); BayesBoost (R/`bnlearn`)	Compare deep vs probabilistic approaches
Configurations	Default hyperparameters	Default batch size/epochs; DECAF ND (no-debias) causal variant	Ensure comparability and reproducibility
Iterations	150 per model per dataset (distinct random seeds)	Synthetic data regenerated each iteration	Average results to reduce random variability
Evaluation metrics	Fidelity	TV/KS Complement, Category/Range Coverage, Boundary Adherence, Correlation (Frobenius norm)	Quantify statistical similarity between real and synthetic data
	Utility	RF, LR, NB classifiers; evaluated via Recall, Precision, F1, AUC-ROC	Assess predictive consistency between real and synthetic data
	Fairness	Accuracy Parity, Predictive Rate Parity, Equalised Odds	Evaluate subgroup equity across sensitive attributes
Hardware/ Software	Experimental environment	Intel Evo CPU (3.8 GHz), 32 GB RAM, RTX 4060 (16 GB); Python 3.12, R 4.3	Execution environment and library versions

those in the real data. TV Complement and KS Complement evaluate similarity in the marginal distributions. Boundary Adherence measures whether synthetic values remain within the observed real-data limits. Higher values across all metrics indicate closer alignment between synthetic and real data.

For categorical variables, BayesBoost, CTGAN, and CopulaGAN successfully captured all subcategories present in the real datasets for both CVD and Diabetes. In contrast, DECAF lagged slightly behind, covering approximately 87% of the subcategories in the CVD dataset. TV Complement scores remained consistently high across all models for the CVD dataset and exceeded 80% for the Diabetes dataset, suggesting that the marginal distributions of categorical variables in the synthetic data closely resembled those of the real data.

Regarding numerical variables, Range Coverage showed a significant decline in the CVD dataset for the deep generative models CopulaGAN, CTGAN, VAE, and DECAF, indicating challenges in capturing the complete range of feature distributions. DECAF also demonstrated poor range coverage in the Diabetes dataset, whereas the other models successfully covered the full numerical range. KS Complement scores exhibited moderate variation, with CTGAN and VAE performing slightly worse than BayesBoost, which achieved the best preservation of the numerical distributions, maintaining KS Complement values above 97%. DECAF recorded the lowest KS Complement scores, with approximately 80% fidelity in the CVD dataset and even lower in the Diabetes dataset, further underscoring its limitations in accurately replicating real data distributions.

Boundary Adherence remained consistently high across all models for both datasets, demonstrating that the synthetic samples adhered well to the expected numerical feature boundaries and no new values were introduced in the generated data.

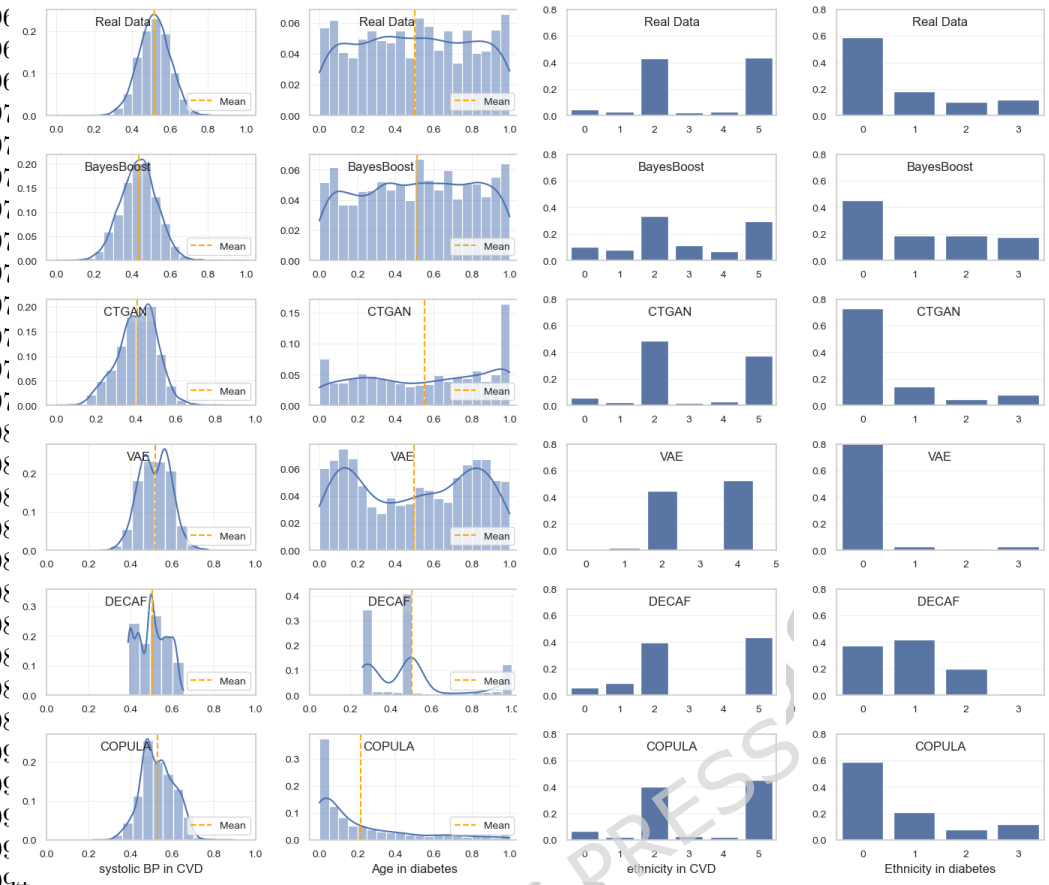
Figure 2 illustrates examples of generated data distributions for selected features, based on randomly chosen synthetic datasets of identical size to the real datasets. For numerical features, such as systolic blood pressure (CVD dataset) and age (Diabetes dataset), BayesBoost and VAE most closely approximated the distributions observed in the real data. In contrast, DECAF exhibited the greatest divergence, generating distributions that noticeably deviated from the original datasets.

Fig. 1: Statistical similarity between synthetic data and real data for Diabetes and CVD datasets. Higher values indicate closer alignment between synthetic and real data.

For categorical features, particularly Ethnicity in both datasets, both DECAF and VAE struggled to accurately replicate the distributions of underrepresented subgroups. DECAF failed to generate any instances for Groups 3 and 4 in the CVD dataset, and for Group 3 in the Diabetes dataset, all of which were already minority subgroups in the original data. Similarly, VAE failed to generate instances for the protected subgroup (Group 2) in both datasets. Furthermore, VAE distorted subgroup proportions, notably over-representing Group 4 in the Diabetes dataset and failing to generate instances for Group 5, the most populous group in the real data.

By contrast, CopulaGAN and CTGAN maintained subgroup distributions more faithfully, with only minor deviations from the real datasets. BayesBoost, however, over-sampled certain minority subgroups, nearly doubling their representation relative to the real data. This behaviour can be attributed to BayesBoost's internal mechanism, which deliberately amplifies minority subgroup instances to enhance class balance during generation.

For the multivariate analysis, Figure 3 presents the average Frobenius norm distances between the correlation matrices of the real and synthetic datasets across



(a) systolic bp from CVD **(b)** age from diabetes **(c)** ethnicity from CVD **(d)** ethnicity from diabetes

Fig. 2: Examples of feature distributions in real and synthetic datasets for selected features

150 iterations for both the CVD and Diabetes datasets. Lower values indicate better preservation of inter-variable relationships, whereas higher values reflect greater divergence from the original data's correlation structure.

In the CVD dataset, BayesBoost achieved the best fidelity, with the lowest distance (0.2385), indicating strong retention of the real data's multivariate dependencies. DECAF (0.3479), CTGAN (0.3787), and CopulaGAN (0.3848) exhibited moderate deviations, while VAE recorded the highest distance (0.7578), suggesting weaker preservation of correlation patterns compared to the other models.

For the Diabetes dataset, discrepancies were more pronounced. BayesBoost again achieved the lowest distance (0.3845), reinforcing its ability to maintain the correlation structure of the real data. In contrast, CTGAN (0.9953) and DECAF (0.9495) displayed substantial deviations, while CopulaGAN (1.4379) and VAE (2.0172) exhibited the highest distances, indicating significant loss of inter-variable relationships within the synthetic data.

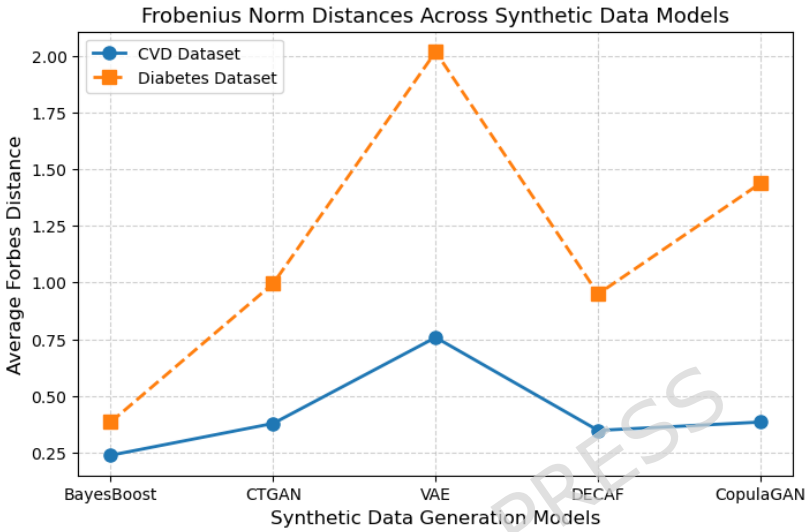


Fig. 3: Average Frobenius Norm Distances across synthetic data generation models

3.1.1 ML Utility

This section presents a comparative analysis of the performance of machine learning classifiers trained on real and synthetic datasets for two use cases: CVD prediction and diabetes prediction. The effectiveness of five synthetic data generation methods—CopulaGAN, CTGAN, VAE, DECAF, and BayesBoost—is evaluated by comparing performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC, against those achieved on real datasets. All performance metrics are summarised in Table 3.

Table 3: Machine Learning Performance Metrics on Real and Synthetic Data for CVD and Diabetes Datasets

			CVD					
M	m[1]	a	b	c	d	e	f
RF	i	0.914 ± 0.000	0.887 ± 0.001	0.889 ± 0.001	0.904 ± 0.001	0.799 ± 0.006	0.919 ± 0.000
	ii	0.588 ± 0.002	0.474 ± 0.007	0.473 ± 0.007	0.597 ± 0.002	0.436 ± 0.006	0.619 ± 0.001
	iii	0.459 ± 0.002	0.408 ± 0.012	0.396 ± 0.011	0.535 ± 0.004	0.558 ± 0.008	0.493 ± 0.001
	iv	0.818 ± 0.002	0.666 ± 0.010	0.687 ± 0.010	0.682 ± 0.004	0.384 ± 0.010	0.829 ± 0.001
	v	0.884 ± 0.001	0.853 ± 0.001	0.854 ± 0.001	0.869 ± 0.001	0.798 ± 0.003	0.896 ± 0.000
LR	i	0.912 ± 0.000	0.889 ± 0.001	0.891 ± 0.001	0.901 ± 0.001	0.873 ± 0.002	0.922 ± 0.000
	ii	0.585 ± 0.002	0.481 ± 0.007	0.479 ± 0.007	0.596 ± 0.002	0.533 ± 0.004	0.642 ± 0.001
	iii	0.463 ± 0.002	0.408 ± 0.012	0.395 ± 0.011	0.552 ± 0.004	0.536 ± 0.005	0.525 ± 0.001
	iv	0.795 ± 0.002	0.680 ± 0.009	0.700 ± 0.009	0.657 ± 0.005	0.548 ± 0.008	0.825 ± 0.001
	v	0.878 ± 0.001	0.864 ± 0.001	0.866 ± 0.001	0.866 ± 0.001	0.849 ± 0.002	0.891 ± 0.000
NB	i	0.763 ± 0.002	0.854 ± 0.002	0.851 ± 0.005	0.845 ± 0.005	0.851 ± 0.005	0.850 ± 0.009
	ii	0.476 ± 0.002	0.542 ± 0.003	0.545 ± 0.003	0.543 ± 0.004	0.491 ± 0.006	0.543 ± 0.006
	iii	0.805 ± 0.003	0.648 ± 0.005	0.653 ± 0.005	0.673 ± 0.004	0.524 ± 0.008	0.590 ± 0.011
	iv	0.340 ± 0.003	0.473 ± 0.005	0.477 ± 0.005	0.463 ± 0.006	0.485 ± 0.009	0.577 ± 0.014
	v	0.868 ± 0.001	0.842 ± 0.002	0.846 ± 0.002	0.830 ± 0.002	0.828 ± 0.003	0.835 ± 0.003
			Diabetes					
M	m	a	b	c	d	e	f
RF	i	0.906 ± 0.001	0.542 ± 0.006	0.545 ± 0.006	0.840 ± 0.002	0.601 ± 0.000	0.882 ± 0.001
	ii	0.873 ± 0.001	0.221 ± 0.018	0.221 ± 0.018	0.786 ± 0.003	0.130 ± 0.000	0.839 ± 0.002
	iii	0.812 ± 0.002	0.281 ± 0.028	0.275 ± 0.027	0.738 ± 0.005	0.151 ± 0.001	0.775 ± 0.003
	iv	0.945 ± 0.001	0.352 ± 0.019	0.382 ± 0.020	0.848 ± 0.004	0.191 ± 0.000	0.916 ± 0.001
	v	0.951 ± 0.001	0.503 ± 0.004	0.500 ± 0.004	0.900 ± 0.001	0.512 ± 0.002	0.947 ± 0.000
LR	i	0.831 ± 0.001	0.539 ± 0.006	0.544 ± 0.006	0.767 ± 0.002	0.489 ± 0.008	0.763 ± 0.001
	ii	0.781 ± 0.002	0.268 ± 0.017	0.255 ± 0.017	0.730 ± 0.002	0.512 ± 0.014	0.653 ± 0.002
	iii	0.757 ± 0.002	0.316 ± 0.026	0.297 ± 0.026	0.784 ± 0.005	0.798 ± 0.027	0.561 ± 0.003
	iv	0.808 ± 0.002	0.406 ± 0.014	0.411 ± 0.012	0.688 ± 0.004	0.421 ± 0.011	0.785 ± 0.002
	v	0.905 ± 0.001	0.505 ± 0.005	0.504 ± 0.005	0.854 ± 0.002	0.640 ± 0.007	0.854 ± 0.001
NB	i	0.790 ± 0.002	0.525 ± 0.005	0.534 ± 0.005	0.517 ± 0.003	0.601 ± 0.000	0.671 ± 0.002
	ii	0.717 ± 0.003	0.379 ± 0.010	0.370 ± 0.008	0.534 ± 0.003	0.140 ± 0.012	0.549 ± 0.002
	iii	0.667 ± 0.004	0.407 ± 0.017	0.377 ± 0.015	0.692 ± 0.007	0.180 ± 0.001	0.502 ± 0.003
	iv	0.777 ± 0.003	0.408 ± 0.006	0.420 ± 0.004	0.439 ± 0.003	0.201 ± 0.003	0.608 ± 0.003
	v	0.875 ± 0.001	0.504 ± 0.005	0.504 ± 0.004	0.564 ± 0.002	0.500 ± 0.000	0.758 ± 0.002

[1]Performance metrics; i: accuracy; ii: f1-score; iii: recall; iv: precision; v: auc-roc;
Training Data; a: real data; b: CopulaGAN; c: CTGAN; d: VAE; e: DECAF; f: BayesBoost

Across both datasets, the RF classifier consistently achieved the highest performance when trained on real data, demonstrating its robustness. For CVD prediction, RF achieved an accuracy of 0.914 ± 0.000 and an AUC-ROC of 0.884 ± 0.001, while for diabetes, the corresponding metrics were 0.906 ± 0.001 and 0.951 ± 0.001. Logistic Regression (LR) also performed strongly on the CVD data, recording an accuracy of 0.912 ± 0.000 and an AUC-ROC of 0.878 ± 0.001. However, its performance declined for diabetes prediction (accuracy of 0.831 ± 0.001 and AUC-ROC of 0.905 ± 0.001)

compared to RF. Naive Bayes (NB) demonstrated the lowest performance across both datasets, achieving an accuracy of 0.763 ± 0.002 for CVD and 0.790 ± 0.002 for diabetes.

Recall varied significantly across classifiers and datasets. For the CVD data, NB achieved the highest recall (0.805 ± 0.003), outperforming RF (0.459 ± 0.002) and LR (0.463 ± 0.002). In contrast, for diabetes prediction, RF achieved the highest recall (0.812 ± 0.002), outperforming LR (0.757 ± 0.002) and NB (0.667 ± 0.004).

For CVD prediction, synthetic data generated by CopulaGAN and CTGAN yielded reasonable performance for RF and LR in terms of accuracy, AUC-ROC, and recall, with only minor deviations from the performance observed on real data. However, for precision and F1-score, larger differences were observed, reaching up to 0.2 in some cases. NB's performance on CopulaGAN- and CTGAN-generated data remained suboptimal, with decreased recall but slight improvements in accuracy and precision, reflecting a trade-off between sensitivity and specificity.

Conversely, for diabetes prediction, CopulaGAN and CTGAN exhibited significant performance deterioration across all classifiers. RF accuracy on CopulaGAN-generated data dropped sharply to 0.542, with AUC-ROC falling to 0.503. Similarly, CTGAN yielded an RF accuracy of 0.545 ± 0.006.

The VAE-generated CVD data demonstrated high utility, closely resembling the predictive capabilities achieved with real data across all classifiers. Specifically, RF trained on VAE-generated data achieved an accuracy of 0.904 ± 0.001 and an AUC-ROC of 0.869 ± 0.001. LR similarly achieved strong results with an accuracy of 0.901 ± 0.001 and an AUC-ROC of 0.866 ± 0.001.

For the diabetes dataset, VAE outperformed CopulaGAN, CTGAN, and DECAF, delivering results closest to real data across most metrics and classifiers. RF demonstrated particularly strong performance on VAE data, achieving an accuracy of 0.840 ±

0.002 and an AUC-ROC of 0.900 ± 0.001. LR also performed well, achieving an accuracy of 0.767 ± 0.002 and an AUC-ROC of 0.854 ± 0.002. These results demonstrate that VAE maintained utility and complexity relatively well for challenging datasets, enabling robust ML model performance.

DECAF, however, demonstrated the weakest performance overall. For CVD, RF and LR accuracies dropped to 0.799 and 0.873, respectively, with corresponding AUC-ROC values of 0.798 and 0.849. Although DECAF slightly improved recall, this was overshadowed by significant reductions in precision and overall model utility. On the diabetes dataset, DECAF's performance collapsed almost entirely, with RF and NB accuracy falling to 0.601 ± 0.000 and metrics such as precision and recall approaching 0.151 and 0.191, indicating a major loss of predictive signal.

Among all synthetic data generation methods, BayesBoost consistently achieved the best performance across most evaluation metrics. For CVD data, RF showed a slight improvement when trained on BayesBoost-generated data, achieving an accuracy of 0.919 compared to 0.914 on real data, and a higher F1-score (0.619 vs 0.588). Similarly, LR trained on BayesBoost data achieved an accuracy of 0.922 and an F1-score of 0.642, outperforming the results achieved with real data (accuracy of 0.912 and F1-score of 0.585). NB also showed noticeable improvements in accuracy (0.850 vs 0.763) and F1-score (0.543 vs 0.476), although recall slightly declined.

For the diabetes dataset, BayesBoost maintained relatively strong predictive utility compared to other synthetic data methods. RF performance remained high, with only a marginal decline in accuracy (0.882 vs 0.906) and AUC-ROC (0.947 vs 0.951). However, LR and NB experienced more pronounced declines: LR's accuracy fell from 0.831 to 0.763, and NB's accuracy decreased from 0.790 to 0.671.

3.1.2 Fairness Analysis

This section presents an evaluation of the fairness of training data by comparing fairness metrics across different classifiers and subgroups defined by the sensitive attribute.

The primary objective is to assess the baseline fairness of the real datasets and to examine the extent to which synthetic data generated by various models preserves or distorts this fairness. Fairness is quantified using three established metrics: Accuracy Parity, Equalised Odds, and Predictive Rate Parity, as detailed in Tables 4 and 5. Real data serves as the benchmark against which the performance of each synthetic data generation model is evaluated.

- **CVD Data**

Accuracy parity: The real data consistently demonstrates fairness across all classifiers (LR, Rf and NB) and ethnic subgroups. The protected subgroup (Group 0 - Asians) exhibits classification accuracy comparable to all other groups (Groups 1–5), with the ratio differences not exceeding 0.04 across classifiers. This finding highlights the capacity of real data to maintain equitable classification performance across diverse subgroups within the sensitive attribute of ethnicity.

Synthetic data generated by all methods largely preserves the accuracy parity observed in the real data for groups 1, 2, 4, and 5, with only minor deviations that can be considered negligible. However, for group 3, the ratio deviation from real data is slightly higher when using the NB classifier, reaching more than 0.1. Despite this, the overall fairness remains well preserved relative to the baseline.

Equalised odds: For models trained on real data, NB achieved the best fairness performance, with deviations from the ideal parity value of 1 remaining within 0.05 across all ethnic groups. LR and RF also maintained reasonable parity overall; however, noticeable deviations emerged, particularly for Group 5, where parity ratios reached 1.19 (LR) and 1.23 (RF), indicating a systematic disadvantage to the protected subgroup (Group 0).

When classifiers were trained on synthetic datasets, it became evident that all generative models struggled to maintain equalised odds parity, particularly for Groups

Table 4: Fairness Metrics for Ethnic Groups Measured Relative to Sensitive Group 0 - CVD Data

		Accuracy Parity			Equalised Odds			PRP		
G[1]	D[2]	LR	NB	RF	LR	NB	RF	LR	NB	RF
1	a	1.01	0.99	1.02	1.06	0.95	1.11	0.81	0.77	0.89
	b	1.06	1.08	1.05	1.63	1.32	1.44	1.08	1.33	1.04
	c	1.06	1.08	1.05	1.70	1.33	1.50	1.10	1.29	1.03
	d	1.07	1.10	1.08	1.84	1.29	1.57	1.08	1.47	1.29
	e	1.08	1.09	1.04	1.81	1.74	1.89	1.36	1.45	1.04
	f	1.06	1.05	1.05	1.66	1.71	1.50	1.01	1.18	1.00
2	a	0.98	0.99	0.99	1.10	1.00	1.16	0.81	1.07	0.86
	b	0.97	0.98	0.97	0.73	0.85	0.79	0.88	0.93	0.88
	c	0.97	0.98	0.98	0.73	0.85	0.79	0.89	0.92	0.89
	d	0.98	1.00	0.98	0.98	0.88	0.92	0.88	1.03	0.92
	e	1.00	1.00	0.98	1.10	1.13	1.20	1.01	1.06	1.02
	f	0.98	0.98	0.98	0.94	1.08	1.04	0.83	0.89	0.83
3	a	0.96	1.01	0.97	0.88	1.02	0.98	0.82	1.20	0.88
	b	0.92	0.89	0.92	0.35	0.69	0.37	0.53	0.66	0.55
	c	0.92	0.89	0.92	0.35	0.65	0.37	0.51	0.63	0.55
	d	0.94	0.93	0.95	0.68	0.72	0.73	0.76	0.79	0.87
	e	0.93	0.94	0.91	0.55	0.52	0.64	0.66	0.68	0.68
	f	0.93	0.87	0.94	0.46	0.55	0.60	0.73	0.44	0.74
4	a	1.00	0.98	1.01	0.97	0.97	1.05	0.70	0.74	0.77
	b	0.99	0.98	1.00	0.96	1.10	1.03	0.67	0.73	0.76
	c	0.99	0.98	1.00	0.93	1.10	0.96	0.67	0.74	0.73
	d	1.01	0.97	1.00	1.41	1.09	1.23	0.79	0.74	0.79
	e	1.03	1.04	0.99	1.45	1.45	1.76	0.95	1.02	0.89
	f	1.02	1.04	1.04	1.39	1.53	1.59	0.81	1.00	0.93
5	a	0.99	0.96	0.99	1.19	0.95	1.23	0.85	0.97	0.89
	b	0.98	0.99	0.99	1.03	1.03	1.05	0.86	0.96	0.88
	c	0.99	0.99	0.99	1.02	1.02	1.04	0.87	0.95	0.88
	d	0.99	1.00	1.00	1.22	1.03	1.16	0.86	1.04	0.99
	e	1.01	1.01	0.99	1.37	1.33	1.41	1.05	1.09	1.05
	f	1.00	1.03	1.00	1.26	1.33	1.29	0.83	0.96	0.82

Fairness metrics are measured relative to Group 0, with values closest to 1 being optimal and highlighted in white, while darker red shades indicate greater deviation from 1.

[1]G: Ethnicity groups; 0: Asian; 1: Black; 2: Mixed; 3: Others; 4: unknown; 5: White

[2]Data; a: real data; b: CopulaGAN; c: CTGAN; d: VAE; e: DECAF; f: BayesBoost

1 and 3. Synthetic data often introduced higher parity values for Group 1 and significantly lower parity values for Group 3, indicating a pronounced amplification of existing biases.

Synthetic datasets produced by CopulaGAN, CTGAN, and DECAF led to substantial fairness distortions. These models caused dramatic parity shifts, particularly in Groups 1, 2 and 3, where parity ratios fell sharply—reaching approximately 50–70%

lower than Group 0 for Group 3, and conversely achieving 40–90% better parity for Group 1 across LR and RF classifiers. These extreme deviations highlight that CopulaGAN, CTGAN, and DECAF severely amplified fairness concerns, disproportionately benefiting or disadvantaging certain groups relative to the protected subgroup. Although fairness discrepancies in Groups 4 and 5 were somewhat less severe, their deviations remained notable compared to the real data baselines and 1.

BayesBoost also exhibited considerable fairness disparities across most groups. While it maintained relatively stable parity for Group 2, significant fairness violations were observed for Groups 1, 3, 4, and 5. In Group 1, parity ratios rose as high as 1.66, 1.71, and 1.50 for LR, NB, and RF respectively, suggesting substantial over-prediction compared to the protected subgroup. In contrast, Group 3 exhibited severe under-prediction, with parity ratios falling between 0.46 and 0.60 across classifiers. Furthermore, fairness deviations in Groups 4 and 5 remained consistently elevated, with parity values exceeding 1.25 in several cases.

VAE, on the other hand, achieved equalised odds values closest to those observed in the real data compared to the other generation methods. While it did introduce some fairness concerns, these deviations were substantially less severe, particularly when combined with NB and RF. This was most evident across Groups 2, 3, 4, and 5, where VAE consistently maintained parity values nearer to those in the baseline.

Predictive rate parity: In the real dataset, the PRP outcomes do not conform to the expectations of ideal fairness. Across all classifiers, models consistently demonstrated better predictive parity for the protected subgroup (Group 0) compared to other ethnic groups. This consistent advantage in prediction rates for Group 0 reflects a fairness imbalance inherently embedded within the original data.

Upon training classifiers on synthetic datasets, pronounced fairness variations emerged across generative models. In particular, all synthetic models reversed the fairness trends observed in real data for Group 1. Whereas Group 0 initially exhibited

higher PRP values, synthetic data disproportionately favoured Group 1. This shift was especially pronounced with the NB classifier, where parity ratios indicated over a 40% advantage for Group 1 relative to Group 0. Although fairness concerns persisted across classifiers, the extent of fairness deviation was notably reduced with RF, where PRP values were shifted closer to 1, thereby rendering the predictions for Group 1 more comparable to those for Group 0. In addition, across all generative models, additional fairness disparities were introduced in Group 3 compared to the real data baseline.

For Groups 2, 4, and 5, CopulaGAN, CTGAN, and VAE largely preserved the fairness structure observed in the real data, without inducing substantial deviations. While DECAF - like other generative models- introduced pronounced fairness distortions, particularly affecting Groups 1 and 3. Nonetheless, in Groups 2, 4, and 5, DECAF marginally improved fairness by narrowing the gap in parity ratios between the protected subgroup and the other groups.

Among the evaluated methods, BayesBoost demonstrated the most consistent preservation of fairness. It either retained the original PRP values or yielded ratios that remained closest to 1 across all groups. Particularly, RF trained on BayesBoost-generated data exhibited highly stable parity, achieving a PRP of 1.00 for Group 1 and 0.74 for Group 3, outperforming all other model-classifier combinations by minimising deviations from 1 or from real data parity. Logistic Regression (LR) combined with BayesBoost similarly demonstrated competitive fairness, albeit with a slight underestimation for Group 3 (PRP = 0.73) compared to real data.

- **Diabetes Data**

Accuracy parity: On the real data, AP results were consistently close to the ideal value of 1 across all ethnic groups and classifiers (LR, NB, RF). Deviations were minimal, with values ranging between 0.97 and 1.04, indicating that classification accuracy was relatively balanced between the protected subgroup (Group 2, Asian ethnicity) and the other ethnic groups.

Table 5: Fairness Metrics for Ethnic Groups Measured Relative to Sensitive Group 2 - Diabetes Dataset

		Accuracy Parity			Equalised Odds			PRP		
G[1]	D[2]	LR	NB	RF	LR	NB	RF	LR	NB	RF
0	a	1.02	1.02	0.99	1.15	1.12	1.01	1.02	1.08	1.00
	b	1.00	1.01	0.98	2.17	1.48	1.47	9.82	1.96	10.27
	c	1.00	1.03	0.99	1.44	1.91	1.07	10.89	3.53	9.49
	d	1.14	1.05	1.04	1.06	1.10	0.98	1.27	1.15	1.16
	e	0.94	0.94	0.94	0.91	0.12	0.15	1.08	0.11	0.13
	f	1.02	1.03	1.04	1.20	1.36	1.21	1.05	1.18	1.02
1	a	1.02	1.04	0.97	1.21	1.22	0.99	1.02	1.10	0.96
	b	1.01	1.00	0.96	1.86	1.34	1.24	8.42	2.59	7.82
	c	0.99	1.02	0.98	1.13	1.44	1.08	8.72	2.94	8.69
	d	1.17	1.11	1.05	1.05	1.14	1.01	1.35	1.26	1.17
	e	0.92	0.92	0.92	0.32	0.11	0.14	1.90	0.20	0.18
	f	1.04	1.13	1.02	1.09	1.53	1.13	1.21	1.42	1.04
3	a	1.01	1.04	0.99	1.13	1.11	0.97	0.98	1.09	0.99
	b	1.03	1.03	1.00	1.75	1.52	1.31	4.90	2.33	5.54
	c	0.99	1.01	1.00	1.04	1.38	1.09	3.98	3.29	4.37
	d	1.06	0.99	1.00	0.86	1.12	0.84	1.17	1.03	1.11
	e	0.99	0.99	0.99	0.91	0.23	0.21	1.01	0.16	0.19
	f	1.00	1.02	0.95	0.91	1.51	0.90	1.08	1.07	0.95

Fairness metrics are measured relative to Group 2, with values closest to 1 being optimal and highlighted in white, while darker red shades indicate greater deviation from 1.

[1]G: Ethnicity groups; 0: Caucasian; 1: African American; 2:Asian; 3:Others.

[2]Data; a: real data; b: CopulaGAN; c: CTGAN; d: VAE; e: DECAF; f: BayesBoost

When synthetic data was used, CopulaGAN, CTGAN, and BayesBoost largely preserved the accuracy parity observed in real data across all groups and classifiers. AP ratios remained between 0.96 and 1.04 for most combinations, effectively mirroring real data performance. Although BayesBoost combined with NB introduced some disparities in group 1, the deviations were not substantial.

VAE exhibited a slightly different trend: while fairness was mostly preserved, it introduced noticeable parity inflation, particularly in Groups 0 and 1. For instance, under VAE, Group 0 with LR reached an AP of 1.14 (compared to 1.02 in real data), while Group 1 showed AP values of 1.17 (LR), 1.11 (NB), and 1.05 (RF). Although these deviations are also moderate, they indicate that VAE may increase classification accuracy disproportionately for some groups, potentially shifting fairness dynamics.

DECAF displayed a slight reduction in AP values across groups, most notably for Groups 0 and 1, where AP dropped to 0.94 and 0.92, respectively. However, the magnitude of these deviations remained below 10%, suggesting that while DECAF introduced a slight disadvantage for these groups relative to Group 2, it did not significantly compromise overall fairness.

Equalised odds: On the real data, EO results demonstrated reasonably good fairness, particularly for RF, where parity values remained close to the ideal value of 1 across all groups (ranging between 0.97 and 1.01). However, larger deviations were observed for LR and NB. Notably, in Group 1, parity ratios reached 1.21 (LR) and 1.22 (NB), indicating that the protected subgroup (Group 2) was systematically disadvantaged relative to Group 1. Similar patterns were seen for Group 0 and Group 3, where parity ratios reached 1.15 (LR) and 1.12 (NB).

When classifiers were trained on synthetic datasets, substantial fairness distortions became apparent, particularly for CopulaGAN and CTGAN. These models introduced severe parity inflations, especially in Group 0 (e.g., CopulaGAN_LR = 2.17, CTGAN_LR = 1.44) and Group 3 (CopulaGAN_LR = 1.75, CopulaGAN_NB = 1.52). Such high ratios suggest significant over-prediction relative to the protected subgroup, thus amplifying existing fairness gaps in the real data.

DECAF exhibited the poorest performance regarding equalised odds. In both Group 0, 1 and 2, DECAF produced near-zero values (between 0.11 and 0.23 with NB and RF), indicating that correct and incorrect prediction rates between the protected group and others became dramatically imbalanced. This critical fairness breakdown renders DECAF unsuitable for fairness-sensitive applications in this context.

Among all models, VAE showed the best preservation of equalised odds. Across Groups 0 and 1, VAE maintained EO ratios close to 1, achieving values of 1.06 (LR), 1.10 (NB), and 0.98 (RF) for Group 0, and 1.05 (LR), 1.14 (NB), and 1.01 (RF) for Group 1. These results represent only minor fairness deviations relative to the real

data baseline. For Group 3, VAE achieved parity values of 0.86 (LR), 1.12 (NB), and 0.84 (RF), which, although slightly lower, still performed substantially better than CopulaGAN, CTGAN, and DECAF.

BayesBoost, by contrast, demonstrated a more mixed fairness profile. While it generally performed better than CopulaGAN, CTGAN, and DECAF, it consistently introduced larger deviations from 1 than VAE. In Group 0, BayesBoost recorded values of 1.20 (LR), 1.36 (NB), and 1.21 (RF), notably inflating the advantage compared to VAE. Particularly concerning was the combination of BayesBoost with NB, where the parity ratio reached 1.36, higher than the baseline real data value of 1.12, indicating a significant worsening of fairness compared to LR and RF.

When comparing LR and RF across VAE and BayesBoost, VAE consistently pushed equalised odds ratios closer to 1 for Groups 0 and 1. RF and LR trained on VAE-synthetic data achieved EO values notably closer to the fairness ideal compared to those trained on BayesBoost-synthetic data. However, for Group 3, a different pattern emerged: VAE lowered parity values to 0.86 (LR) and 0.84 (RF), slightly below the ideal, whereas BayesBoost combined with LR achieved a parity value closer to 1 for this group. This suggests that for preserving equalised odds parity in Group 3, BayesBoost combined with LR slightly outperformed VAE. Nevertheless, overall, VAE combined with RF achieved the best overall equalised odds parity across all groups, with BayesBoost combined with LR providing a complementary solution for achieving more balanced fairness.

Predictive Rate parity: On the real data, PRP values across LR, NB, and RF classifiers were consistently close to the ideal value of 1 for all ethnic groups, suggesting reasonably good fairness. Group 0 exhibited PRP values of 1.02 (LR), 1.08 (NB), and 1.00 (RF), while Group 1 ranged between 0.96 and 1.10, and Group 3 displayed similarly balanced values between 0.98 and 1.09. These results indicate a slight systematic advantage for Group 0 that does not exceed 2% but without severe deviations from ideal fairness.

When classifiers were trained on synthetic datasets, significant fairness disparities emerged depending on the generative model. Similar to the Equalised Odds findings, CopulaGAN, CTGAN, and DECAF introduced major fairness violations, substantially inflating PRP values in Groups 0, 1, and 3, often exceeding a ratio of 2 across all classifiers. These extreme inflations reflect severe fairness issues, particularly as the models performed markedly worse on the protected Group 2. Moreover, DECAF caused PRP values to collapse entirely to near zero for NB and RF classifiers, indicating a complete breakdown in fairness for these groups.

In contrast, VAE demonstrated considerably greater stability. Although over-predictions persisted, PRP values remained moderately close to 1 across all groups and classifiers. For instance, in Group 0, VAE achieved PRP values of 1.27 (LR), 1.15 (NB), and 1.16 (RF), and for Group 3, ratios remained within the range of 1.11–1.17. These results suggest that VAE effectively mitigated the extreme fairness distortions observed with other deep generative models while maintaining relatively stable predictive parity compared to other models.

BayesBoost achieved the best preservation of predictive rate parity among all generative models, particularly when paired with LR and RF classifiers. PRP values remained consistently close to 1 across groups, with Group 0 reaching 1.05 (LR) and 1.02 (RF), and Group 3 achieving 1.08 (LR) and 0.95 (RF). BayesBoost outperformed VAE in several scenarios, especially when used with Random Forest, maintaining predictive parity closest to both the real data baseline and the fairness ideal.

Overall, the results indicate that BayesBoost, particularly when paired with RF, offers the most robust approach for maintaining fairness in predictive rate parity. In contrast, VAE combined with RF and NB emerged as the most effective option for preserving fairness in equalised odds. Conversely, CopulaGAN, CTGAN, and DECAF require cautious consideration before use in fairness-critical healthcare applications. These findings highlight the importance of rigorous fairness evaluation when selecting

synthetic data generation techniques for machine learning models involving protected attributes.

4 Discussion

This study presents a comprehensive evaluation of five prominent synthetic data generation models applied to healthcare tabular datasets, assessing their fidelity, machine learning utility, and fairness preservation. While synthetic data is widely promoted as a solution to privacy concerns and data scarcity [3, 31], our results highlight that many models still face substantial challenges in producing clinically reliable, fair, and structurally faithful synthetic datasets—especially when sensitive attributes such as ethnicity are involved.

After assessing data fidelity and structural preservation, BayesBoost emerged as the most robust method overall. It achieved the highest statistical fidelity, with the lowest Frobenius norm distances and superior boundary and range coverage. Its probabilistic boosting mechanism improved the representation of minority subgroups, but we caution that overboosting, while improving minority representation, can also distort the underlying data distribution and inadvertently introduce new forms of bias or alter the structural integrity of the dataset yielding to the butterfly effect [32]. Careful calibration is therefore necessary when applying such techniques, particularly in fairness-sensitive domains. These results support recent work showing the effectiveness of probabilistic approaches in tabular domains with complex feature types [33].

In contrast, deep generative models—particularly TVAE and DECAF—struggled to preserve feature interdependencies. Although TVAE produced reasonably realistic marginal distributions, it exhibited the highest multivariate deviations in both datasets, highlighting limitations in capturing complex relational structures. This fidelity degradation is particularly concerning given the relatively modest size of the CVD dataset (10,000 instances), especially considering that in healthcare contexts,

larger datasets are often unattainable due to strict privacy regulations or limited resources hence the need to data generation. One might expect that with larger datasets, where overfitting risks are reduced, models would generalise relational structures more effectively. However, our findings indicate otherwise: even at these moderate sample sizes, deep generative models failed to robustly reconstruct multivariate dependencies. Prior studies have suggested that generative models require substantially larger and more balanced datasets to adequately capture multimodal feature interactions [14]. Yet, our results show that differences in dataset size between CVD and diabetes data (10,000 vs 1,879 instances) did not significantly influence fidelity outcomes, suggesting that model architecture and learning dynamics, rather than dataset size alone, critically limit fidelity in healthcare tabular data synthesis.

In predictive utility, BayesBoost again led performance, matching or exceeding real data benchmarks in some cases, particularly in CVD prediction. This means that simpler models can outperform deep networks when diversity and structure preservation matter. VAE showed the best performance among deep models, particularly for diabetes, but CTGAN and CopulaGAN exhibited severe degradation, especially for AUC-ROC and accuracy where classifier performance collapsed to near-random levels.This significant deterioration likely stems from the inability of GAN-based models to effectively handle the sparsity, heterogeneity, and small sample subgroup distributions characteristic of tabular healthcare data without extensive tuning [11].

Classifier choice also influenced results. RF consistently achieved strong and balanced performance across both datasets, real and synthetic, reinforcing its suitability for healthcare tasks involving complex, heterogeneous data types. In contrast, NB underperformed overall but occasionally delivered higher recall; highlighting the importance of selecting classifiers based on clinical priorities such as sensitivity versus specificity. This observation underscores the need for a context-driven model evaluation approach rather than a reliance on a single global performance metric.

Fairness, however, remains the most pressing concern since algorithmic decisions can directly influence patient outcome. Our findings reveal that fairness imbalances already exist in the real datasets—particularly for equalised odds and PRP—. Rather than mitigating these imbalances, generative models exacerbated them. Moreover, across both datasets, all generative models introduced extreme fairness violations primarily affecting the most underrepresented groups in the real data. These minority groups, suffered the largest distortions in fairness metrics post data generation. This systematic deterioration highlights that models not only failed to address existing disparities but actively amplified them—a failure that could seriously undermine equity in clinical AI systems if unchecked.

CopulaGAN, CTGAN, and DECAF consistently introduced severe fairness violations across both datasets. In the diabetes data, PRP and equalised odds ratios inflated well above 2 or collapsed to near-zero. DECAF performed worst, producing near zero predictions for some subgroups, even among majority populations. Such extreme distortions likely stem from these models' failure to adequately reproduce the full distribution of sensitive subgroups within the generated data.

BayesBoost exhibited strong and consistent fairness preservation, particularly in predictive rate parity. When paired with RF classifiers, BayesBoost maintained PRP values remarkably close to 1 across all ethnic groups, often improving upon the fairness observed in real data. However, its performance on equalised odds was less stable—particularly when combined with Naive Bayes—resulting in inflated parity ratios for certain subgroups. These findings suggest that while BayesBoost is highly effective in correcting representation bias and maintaining overall class balance and preserving data structure, it still struggle with fairness metrics tied to subgroup-specific decision thresholds, which require finer calibration.

VAE, while weaker in utility and structural fidelity, preserved EO best across all generative models. Its architecture, based on minimising instance-level reconstruction

loss rather than adversarial optimisation, may inherently support more equitable feature distributions. This aligns with previous work highlighting the relative fairness stability of VAEs in healthcare settings [].

These results reinforce the need for synthetic data pipelines to integrate robust, multi-dimensional evaluation frameworks. Utility or fidelity alone is insufficient: fairness must be explicitly assessed, audited, and optimised. Model–classifier interactions should be systematically tested, as fairness and utility outcomes can shift dramatically depending on pairing. Above all, synthetic data systems must disclose known limitations —including subgroup risks— and define safe use cases where fairness or fidelity thresholds are met.

4.1 Towards Robust and Accountable Synthetic Data Generation in Healthcare

The findings of this study underscore the urgent need for a structured, multi-dimensional framework when developing machine learning-based synthetic data generation systems—particularly in fairness-critical domains such as healthcare. Rather than relying on a single generative approach or isolated metric, we advocate for an iterative, modular pipeline that integrates multiple synthetic data generation strategies—such as probabilistic models, GAN-based methods, and autoencoders—and evaluates them holistically across three core dimensions: statistical fidelity, predictive utility, and fairness.

Crucially, this framework must be cyclic and adaptive. If a particular generative model fails to meet acceptable thresholds for fidelity or fairness, it should trigger a revision of the generation process—be it through reparameterisation, retraining, or the integration of fairness-aware learning objectives. Furthermore, to capture the full range of model behaviour, multiple classification models (e.g. RF, LR, NB) should be paired systematically with each synthetic data generator. This is essential to

identify performance variations that may arise due to interactions between the inductive biases of classifiers and the structural properties of the synthetic data. Certain model–generator combinations may obscure or amplify disparities, and only by evaluating them together can we ensure robust, generalisable conclusions about the utility and fairness of synthetic datasets.

While current evaluation practices largely focus on statistical similarity and classifier performance, they often lack explicit thresholds for determining when fairness violations become unacceptable or clinically dangerous. Our findings reveal that even models preserving overall fidelity or utility may introduce extreme subgroup disparities in precision, recall, or decision thresholds. To address this, the field needs standardised, context-aware fairness deviation metrics—such as acceptable tolerance bands around a baseline (e.g., parity ratios between 0.9 and 1.1)—tailored to specific applications and aligned with ethical guidelines. This would enable developers and regulators to distinguish between benign and harmful deviations, particularly in high-stakes settings like diagnostics or risk stratification.

Beyond traditional metrics, we also recommend incorporating causal structure validation into synthetic data evaluation. While existing fairness metrics assess statistical parity, they may miss latent structural shifts that alter the interpretability or utility of synthetic data. Approaches such as structural causal models (SCMs) or conditional independence testing can help assess whether synthetic data preserves the underlying causal pathways of real-world variables—essential for ensuring synthetic data supports safe and valid inference [34, 35].

Finally, we stress that synthetic data systems must aim to address both representation bias (e.g. over- or under-generation of sensitive subgroups) and algorithmic performance fairness. Our results suggest that most current generative models struggle to satisfy both simultaneously, with some amplifying subgroup disparities. Consequently, when a model cannot meet both objectives, its limitations must be explicitly

documented—including subgroup-specific fairness risks and clearly defined safe use cases. Transparent reporting of synthetic data limitations should be treated as essential, not optional, particularly in regulated domains such as healthcare and finance.

5 Conclusion

This study demonstrates that while synthetic data generation holds significant promise for healthcare applications, achieving robust fidelity, predictive utility, and fairness remains a substantial challenge. By evaluating five representative models—BayesBoost, CTGAN, TVAE, CopulaGAN, and DECAF—on two healthcare datasets, we provided an integrated comparison across statistical, predictive, and fairness dimensions.

Among the evaluated methods, BayesBoost, particularly when combined with Random Forest, consistently delivered the most reliable outcomes across all dimensions, preserving clinical validity while mitigating subgroup disparities. VAE also emerged as a valuable alternative among deep generative models, achieving superior fairness preservation, particularly for equalised odds, albeit at some cost to overall utility.

Conversely, GAN-based models frequently amplified existing biases and failed to replicate critical multivariate dependencies, raising serious concerns regarding their deployment in fairness-sensitive healthcare settings. The pronounced fairness violations among underrepresented groups underscore that fidelity and predictive accuracy alone are insufficient markers of synthetic data quality.

Overall, the strength of this work lies in its unified, multi-criteria evaluation framework that jointly examines fidelity, utility, and fairness—dimensions often studied in isolation. The findings highlight the importance of incorporating fairness auditing as a core, mandatory component of synthetic data pipelines, evaluated alongside traditional utility and fidelity metrics. The proposed multi-criteria framework provides

a practical foundation for the responsible selection and benchmarking of synthetic data models for healthcare applications, emphasising the importance of structured evaluation frameworks and transparency regarding model limitations as prerequisites for safe and ethical deployment. Until such frameworks are standardised and rigorously applied, the clinical use of synthetic data—particularly for fairness-sensitive applications—should proceed only under strict regulatory oversight and with explicit subgroup-specific validation.

5.1 Limitations and Future Work

This study assumes that the datasets used (CVD and Diabetes) are representative of broader healthcare data structures and that model performance on these datasets can generalise to similar tabular healthcare domains. Although the evaluation covers key indicators of fidelity, predictive utility, and fairness, it remains limited to two datasets and a selection of five established models. The analysis does not include emerging generative paradigms such as diffusion or transformer-based models, nor does it account for longitudinal or multi-modal data, as it focuses exclusively on tabular structures.

Future work should extend these evaluations to larger and more heterogeneous datasets and explore recent generative approaches such as diffusion and transformer-based architectures to further validate and generalise the findings. Additionally, translating the proposed framework into practical bias-auditing and model certification tools could support its application in clinical trial simulation, digital health research, and regulatory evaluation of synthetic data systems.

Acknowledgements.

Declarations

- Funding: This work was funded by the Regulators Pioneer Fund, Department for Science, Innovation and Technology. This work was also supported by the UK Regulatory Science and Innovation Networks – Implementation Phase: Human Health CERSIs programme through the project RADIANT: Regulatory Science Empowering Innovation in Transformative Digital Health and AI (Grant Ref: MCPC24031), funded by the Medical Research Council (MRC) and Innovate UK.

- Conflict of interest/Competing interests : Not applicable
- Ethics approval and consent to participate For the use of CPRD data to generate synthetic data: this was covered by CPRD's Database Research Ethics Approval (IRAS number: 242149)
- Consent for publication: Not applicable
- Data availability: CPRD cardiovascular disease synthetic dataset used in this paper can be requested from CPRD (https://cprd.com/cprd-cardiovascular-disease-synthetic-dataset). The diabetes dataset is publicly available on Kaggle (https://www.kaggle.com/datasets/rabieelkharoua/diabetes-health-dataset-analysis)
- Materials availability: Not applicable
- Code availability : Not applicable
- Author contribution: P.M. and R.B. initiated the project and conceived the overall study design. A.T. supervised the project and experiments, providing critical input on result interpretation and the discussion of their implications. D.A. designed the methodology, carried out the experiments, performed the data analysis, and served as the lead author of the manuscript. B.D contributed expertise on fairness metrics and supported the implementation of the BayesBoost model. All authors contributed to revising the manuscript and approved the final version for submission.

References

[1] Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., al.: Synthetic Data – what, why and how? Preprint at arXiv:2205.03257 (2022). https://doi.org/10.48550/arXiv.2205.03257

[2] Quintana, D.S.: A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. Elife **9**, 53275 (2020)

[3] Hernandez, M., Epelde, G., Alberdi, A., Cilla, R., Rankin, D.: Synthetic data generation for tabular health records: A systematic review. Neurocomputing **493**, 28–45 (2022)

[4] Harper, P.R., Moore, J.W., Woolley, T.E.: Covid-19 transmission modelling of students returning home from university. Health Systems **10**(1), 31–40 (2021)

[5] Caramelo, F., Ferreira, N., Oliveiros, B.: Estimation of risk factors for covid-19 mortality-preliminary results. MedRxiv, 2020–02 (2020)

[6] Wang, Z., Myles, P., Tucker, A.: Generating and evaluating synthetic uk primary care data: preserving data utility & patient privacy. In: 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS), pp. 126–131 (2019). IEEE

[7] Wang, Z., Myles, P., Tucker, A.: Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. Computational Intelligence **37**(2), 819–851 (2021)

[8] Draghi, B., Wang, Z., Myles, P., Tucker, A.: Bayesboost: Identifying and handling bias using synthetic data generators. In: Third International Workshop on Learning with Imbalanced Domains: Theory and Applications, pp. 49–62 (2021). PMLR

[9] Yadav, P., Gaur, M., Madhukar, R.K., Verma, G., Kumar, P.: Rigorous experimental analysis of tabular data generated using tvae and ctgan. International Journal of Advanced Computer Science & Applications **15**(4) (2024)

[10] Marecha, P., Ye, L.: Generation and evaluation of tabular data in different domains using gans. Asian Journal of Research in Computer Science **16**, 15–27 (2023) https://doi.org/10.9734/ajrcos/2023/v16i1331

[11] Miletic, M., Sariyar, M.: Challenges of using synthetic data generation methods for tabular microdata. Applied Sciences **14**(14), 5975 (2024)

[12] Van Breugel, B., Kyono, T., Berrevoets, J., Schaar, M.: Decaf: Generating fair synthetic data using causally-aware generative networks. Advances in Neural Information Processing Systems **34**, 22221–22233 (2021)

[13] Liu, T., Qian, Z., Berrevoets, J., Schaar, M.: Goggle: Generative modelling for tabular data by learning relational structure. In: The Eleventh International Conference on Learning Representations (2023)

[14] Wang, A.X., Chukova, S.S., Simpson, C.R., Nguyen, B.P.: Challenges and opportunities of generative models on tabular data. Applied Soft Computing, 112223 (2024)

[15] Stoian, M.C., Dyrmishi, S., Cordy, M., Lukasiewicz, T., Giunchiglia, E.: How realistic is your synthetic data? constraining deep generative models for tabular data. arXiv preprint arXiv:2402.04823 (2024)

[16] Nik, A.H.Z., Riegler, M.A., Halvorsen, P., Storås, A.M.: Generation of synthetic tabular healthcare data using generative adversarial networks. In: International Conference on Multimedia Modeling, pp. 434–446 (2023). Springer

[17] Pezoulas, V.C., Zaridis, D.I., Mylona, E., Androutsos, C., Apostolidis, K., Tachos, N.S., Fotiadis, D.I.: Synthetic data generation methods in healthcare: A review on open-source tools and methods. Computational and Structural Biotechnology Journal (2024)

[18] Zand, R., Abedi, V., Hontecillas, R., Lu, P., Noorbakhsh-Sabet, N., Verma, M., Leber, A., Tubau-Juni, N., Bassaganya-Riera, J.: Development of synthetic patient populations and in silico clinical trials. Accelerated Path to Cures, 57–77 (2018)

[19] Pappalardo, F., Russo, G., Tshinanu, F.M., Viceconti, M.: In silico clinical trials: concepts and early adoptions. Briefings in bioinformatics **20**(5), 1699–1708 (2019)

[20] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

[21] Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional gan. Advances in neural information processing systems **32** (2019)

[22] Shahul Hameed, M.A., Qureshi, A.M., Kaushik, A.: Bias mitigation via synthetic data generation: A review. Electronics (2079-9292) **13**(19) (2024)

[23] Jain, A., Brooks, J.R., Alford, C.C., Chang, C.S., Mueller, N.M., Umscheid, C.A., Bierman, A.S.: Awareness of racial and ethnic bias and potential solutions to address bias with use of health care algorithms. In: JAMA Health Forum, vol. 4, pp. 231197–231197 (2023). American Medical Association

[24] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer

Science Conference, pp. 214–226 (2012)

[25] Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data **5**(2), 153–163 (2017)

[26] Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)

[27] Verma, S., Rubin, J.: Fairness definitions explained. In: Proceedings of the International Workshop on Software Fairness, pp. 1–7 (2018)

[28] Datalink, C.P.R.: CPRD cardiovascular disease synthetic dataset (Version 2020.06.001) [Data set]. Clinical Practice Research Datalink (2020). https://doi.org/10.11581/YK6N-B652 . https://doi.org/10.11581/YK6N-B652

[29] kharoua, R.E.: Diabetes Health Dataset Analysis. Kaggle (2024). https://doi.org/10.34740/KAGGLE/DSV/8665939 . https://www.kaggle.com/dsv/8665939

[30] Fusion, B.: GeNIe Modeller. Bayes Fusion (2025). https://www.bayesfusion.com/

[31] Alattal, D.R., Wang, Z., Myles, P., Tucker, A.: Creating synthetic geospatial patient data to mimic real data whilst preserving privacy: *2022 35th international symposium on computer-based medical systems (cbms). In: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), pp. 7–12 (2023). https://doi.org/10.1109/CBMS58004.2023.00183

[32] Ferrara, E.: The butterfly effect in artificial intelligence systems: Implications for ai bias and fairness. Machine Learning with Applications **15**, 100525 (2024)

[33] Draghi, B., Wang, Z., Myles, P., Tucker, A.: Identifying and handling data bias within primary healthcare data using synthetic data generators. Heliyon **10**(2) (2024)

[34] Makhlouf, K., Zhioua, S., Palamidessi, C.: When causality meets fairness: A survey. Journal of Logical and Algebraic Methods in Programming, 101000 (2024)

[35] Binkyte-Sadauskiene, R., Makhlouf, K., Pinzón, C., Zhioua, S., Palamidessi, C.: Causal discovery for fairness. CoRR abs/2206.06685 (2022)