algorithms

MDPI

Article

Iterative Application of UMAP-Based Algorithms for Fully Synthetic Healthcare Tabular Data Generation

Carla Lázaro [1],* and Cecilio Angulo [1,2],*

[1] Intelligent Data Science and Artificial Intelligence Research Center, Technical University of Catalonia, Nexus II Building, Jordi Girona 29, 08034 Barcelona, Spain
[2] Institute of Robotics and Industrial Informatics (CSIC-UPC), Llorens i Artigas 4, 08028 Barcelona, Spain
* Correspondence: carla.lazaro@upc.edu (C.L.); cecilio.angulo@upc.edu (C.A.)

Abstract: Building on a previously developed partially synthetic data generation algorithm utilizing data visualization techniques, this study extends the novel algorithm to generate fully synthetic tabular healthcare data. In this enhanced form, the algorithm serves as an alternative to conventional methods based on Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs). By iteratively applying the original methodology, the adapted algorithm employs UMAP (Uniform Manifold Approximation and Projection), a dimensionality reduction technique, to validate generated samples through low-dimensional clustering. This approach has been successfully applied to three healthcare domains: prostate cancer, breast cancer, and cardiovascular disease. The generated synthetic data have been rigorously evaluated for fidelity and utility. Results show that the UMAP-based algorithm outperforms GAN- and VAE-based generation methods across different scenarios. In fidelity assessments, it achieved smaller maximum distances between the cumulative distribution functions of real and synthetic data for different attributes. In utility evaluations, the UMAP-based synthetic datasets enhanced machine learning model performance, particularly in classification tasks. In conclusion, this method represents a robust solution for generating secure, high-quality synthetic healthcare data, effectively addressing data scarcity challenges.

Keywords: fully synthetic data; UMAP; healthcare tabular data; data augmentation

check for updates

Citation: Lázaro, C.; Angulo, C. Iterative Application of UMAP-Based Algorithms for Fully Synthetic Healthcare Tabular Data Generation. *Algorithms* **2024**, *17*, 591. https://doi.org/10.3390/a17120591

Academic Editor: Ayan Chatterjee

Received: 31 October 2024
Revised: 18 December 2024
Accepted: 19 December 2024
Published: 21 December 2024

1. Introduction

Generative Artificial Intelligence (AI) has recently emerged as a transformative tool for creating new data content—such as text, images, or audio—by modeling patterns observed in training data [1]. Unlike discriminative models, which are typically used for classification and prediction [2], generative models aim to understand the underlying data distribution, enabling them to generate data points that resemble real-world data. This generative characteristic has led to broad practical applications, with one notable area being synthetic data generation, particularly valuable in fields requiring substantial data volume and strict privacy measures.

Additionally, with the exponential growth of machine learning applications, there exists increasing dependency on high-quality and high-volume data. Data-driven solutions across many fields rely on rich datasets for effective model training and analysis, yet data limitations frequently hinder progress. In healthcare, for instance, quality data are crucial for solutions in areas such as disease analysis to triage, diagnosis, or treatment [3–7]. However, data quality issues often stem from limited measurements, while data scarcity and privacy concerns arise due to the sensitive nature of medical records and diagnostic information. To address these issues about data privacy, scarcity, and quality, synthetic data have become a powerful tool.

In the medical field, the potential of synthetic data to improve real-world healthcare challenges such as diagnostics, research, and treatment planning is significant [8–12]. For instance, synthetic data can be deployed to train machine learning models for early cancer detection and prognosis without exposing sensitive patient information. This approach is critical in institutions with limited access to comprehensive datasets. Additionally, synthetic data can simulate patient populations for clinical trials, enabling researchers to evaluate treatment efficacy before conducting expensive and time-intensive trials. This could support the detection of health patterns, disease progression predictions, and, ultimately, clinical decision-making [13]. However, because of the sensitive nature of medical data, taking care of data privacy in the AI field remains essential.

In healthcare and other fields involving sensitive information, synthetic data offers a way to overcome challenges of limited availability, privacy restrictions, and class imbalance, all of which can limit the model performance. Synthetic data allow organizations to improve AI models while remaining compliant with privacy regulations by producing large, diverse datasets that reflect the statistical characteristics of real data without exposing personal information. Analysts from organizations like Gartner forecast that synthetic data could surpass real data in AI applications by 2030, as privacy concerns escalate [14]. Therefore, synthetic data are expected to become a critical component in sectors that are highly regulated or data-sensitive.

Healthcare data generation also encompasses various data types; different approaches have been developed for synthetic data generation, notably Variational Autoencoders (VAEs) [15], Generative Adversarial Networks (GANs) [16], diffusion models [17,18], and normalizing flows [19,20]. These techniques have been extensively studied in the context of healthcare, where data security is paramount, as highlighted in recent reviews [21–23].

While synthetic data generation for medical images has yielded promising results, our focus here is on tabular data generation, which is critical for medical applications but often underrepresented in generative AI studies. There are three primary approaches to generating synthetic tabular data: partially synthetic data, fully synthetic data, and imputation-based methods (see Figure 1). Partially synthetic methods produce datasets that integrate both real data and generated values, providing privacy benefits while preserving data utility. Fully synthetic methods, on the other hand, do not retain any real data, potentially offering enhanced privacy protection. Lastly, imputation methods focus on filling missing values by leveraging available data, typically resulting in datasets that mix real and imputed values. Imputation may utilize only the incomplete dataset or, alternatively, include a complete reference dataset to enhance the generated values.

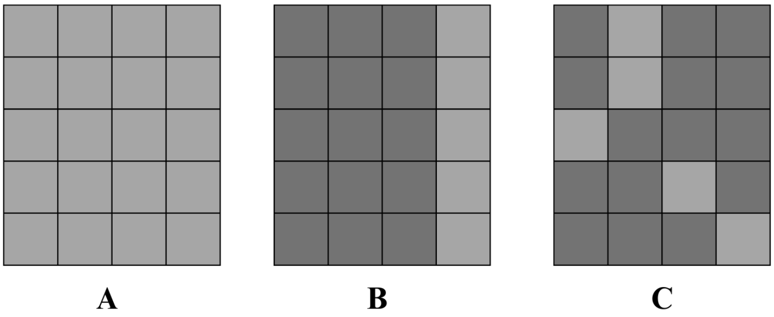


Figure 1. (**A**) Fully synthetic data. (**B**) Partially synthetic data. (**C**) Imputed data. Synthetic values in light gray and real values in dark gray.

Building on the concept of partially synthetic data, a recent study has proposed algorithms that use visualization tools, in particular UMAP, to validate and enhance synthetic data generation [24]. This approach trains a UMAP model on reference data and then generates synthetic samples for one or various attributes. A range of synthetic values are assigned to the attribute that is being generated, and the resulting samples are transformed into the UMAP space. Therefore, the generated values are validated by comparing the UMAP transformation of reference data with that of synthetic samples.

However, while partially synthetic data reduce certain privacy risks, they may still retain identifying characteristics of the original data. This study seeks to expand upon prior work by adapting the partially synthetic generation method for fully synthetic data, which does not contain any real data elements, thereby ensuring a higher level of privacy protection while maintaining data utility.

The primary aim of this paper is to adapt our previously proposed algorithm, initially designed for partially synthetic data generation, to enable the generation of fully synthetic data. This adaptation allows for the creation of datasets that contain no real data components, thus offering an elevated level of privacy protection. By iteratively applying the generation process across all features, the algorithm ensures that the generated dataset is entirely synthetic, expanding its potential applicability.

The structure of this paper is as follows. Section 2 outlines the methodological adaptations for fully synthetic data generation, detailing algorithmic steps, datasets used, evaluation metrics, and potential ethical considerations. Section 3 presents the experimental results obtained by applying the algorithm to a healthcare dataset, alongside a comprehensive quality assessment of the synthetic data. Finally, Section 4 discusses the strengths, limitations, and potential directions for future research.

2. Materials and Methods

The proposed algorithm builds on a previously established methodology for partially synthetic data generation, as introduced in [24]. This initial approach uses data visualization tools, specifically UMAP, to generate synthetic data that maintain essential patterns of the original data. The partially synthetic data generation process relies on two datasets: a *reference dataset* and an *incomplete dataset*. The reference dataset, which contains all the features of interest, serves as a template to generate synthetic values in the incomplete dataset, which contains missing attributes.

In this approach, UMAP is first trained on the reference dataset to create a low-dimensional representation that captures the essential structural characteristics of the data. Synthetic values are then assigned to missing attributes in the incomplete dataset, resulting in "trial" data points that are projected into the same UMAP-reduced space. By comparing these generated samples with the reference data in this low-dimensional space, the samples are validated in two steps.

One validation step calculates whether the generated sample is within a certain radius around a class centroid. In case it is, the sample is "cluster-validated" and the corresponding class or target attribute is also assigned to the sample. The other step involves assigning a reliability score to the synthetic value of the sample. First, the mean value of the synthetic feature of the nearest neighbors is computed. The difference between the mean neighbor value and the synthetic value (feature disparity) determines the reliability score, as indicated in Table 1. Finally, only samples that exceed a certain reliability score threshold are included in the resulting *partially synthetic dataset*.

Table 1. Reliability scores. Value ε depends on the range of the synthetic feature.

Feature Disparity (f_{disp})	**Reliability Scores (r_c)**
$f_{\text{disp}} \leq \varepsilon$	1
$\varepsilon < f_{\text{disp}} \leq 2\varepsilon$	0.9
$2\varepsilon < f_{\text{disp}} \leq 3\varepsilon$	0.8
$3\varepsilon < f_{\text{disp}} \leq 4\varepsilon$	0.7
$4\varepsilon < f_{\text{disp}} \leq 5\varepsilon$	0.6
$5\varepsilon < f_{\text{disp}} \leq 6\varepsilon$	0.5

To generate fully synthetic data, the original methodology is adapted to operate iteratively, thus allowing for the creation of synthetic values for all features sequentially. The setup remains largely consistent with the partially synthetic framework: a reference dataset containing all necessary attributes and a secondary dataset that serves as the incomplete dataset. For fully synthetic data generation, however, the process involves an iterative feature-generation approach, as detailed in Figure 2:

1. Initial setup: the process starts by removing a single attribute from the "incomplete" dataset, thereby creating the initial missing feature scenario. The UMAP-based methodology is then applied as in the original partially synthetic method, generating synthetic values for the removed attribute.
2. Sequential iterations: after completing the first iteration and validating synthetic samples for the initially removed feature, a different attribute is removed from the resulting dataset, and the process is repeated. At each step, the synthetic values generated in prior iterations are preserved.
3. Final outcome: this iterative process is repeated for each attribute in the reference dataset until all features have been synthetically generated. The final result is a fully synthetic dataset that replicates the reference dataset's structure without including any original data entries.

This iterative feature-by-feature synthesis allows for the gradual creation of a fully synthetic dataset that preserves the statistical properties and feature relationships inherent to the reference dataset.

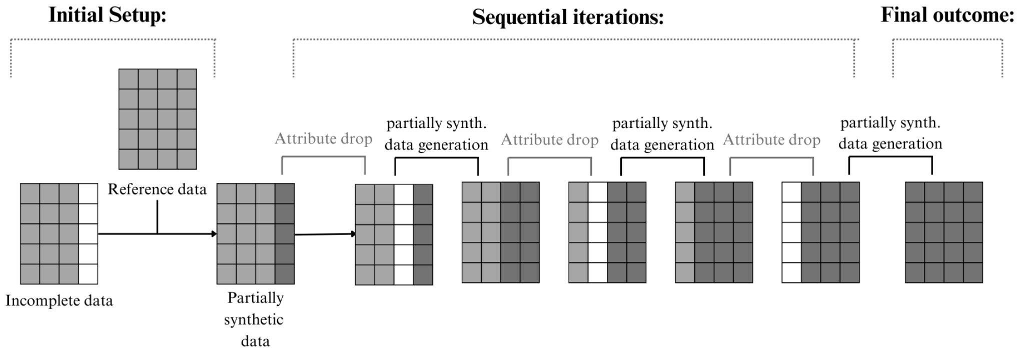


Figure 2. Algorithm workflow.

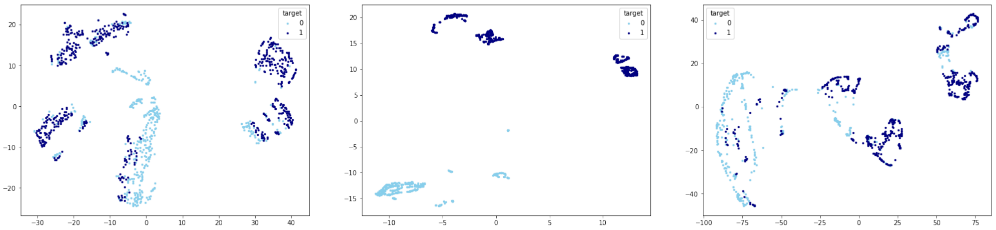
In the following subsections, the algorithm will be described in detail, as well as the dataset used.

2.1. UMAP Transformation and Feature Distribution

UMAP inherently clusters similar data points, ideally reflecting shared attribute values in the low-dimensional space. However, its dependence on initialization might not correctly cluster the samples. Hence, since the methodology is based on the UMAP transformation for the validation, a preliminary visualization of the data after UMAP transformation is crucial to confirm a correct initialization. This starting check will play a vital role in the synthetic data validation phase, as generated data points are validated based on their alignment with neighboring points in the UMAP-reduced space.

This methodology uses UMAP as a dimensionality reduction tool to extend the capabilities of a previously proposed partially synthetic data generation method [24], which is already based on UMAP. Alternative dimensionality reduction techniques, such as t-distributed stochastic neighbor embedding (t-SNE) and TriMap, could also be considered for adaptation in future studies. In the original methodology, these alternative techniques were tested on prostate and breast cancer datasets, as detailed in [24]. The exploration of other dimensionality reduction tools could open interesting research directions, especially for applications in different data domains or structures.

For now, the focus remains on adapting the UMAP-based partially synthetic methodology to fully synthetic data generation. However, for the currently used datasets, UMAP is still the best choice. To illustrate this, Figure 3 compares the transformations achieved by UMAP, t-SNE, and TriMap for cardiovascular disease data. Of the three, UMAP provides the clearest distinction between clusters, further supporting its suitability for the proposed method.



(**a**) t-SNE (**b**) UMAP (**c**) TriMap

Figure 3. Cardiovascular data (Ind-CardioDB) visualizations using different dimensionality reduction techniques.

2.2. Dataset Partitioning and Setup

The initial stage of the methodology involves data preparation, tailored to the nature of the available data. For this experiment, we consider two scenarios. On one side, we assume that an original complete dataset containing all relevant features is accessible. This complete dataset is divided into a reference dataset and an incomplete dataset, simulating missing data for iterative generation purposes. On the other side, if two different datasets are available, one can be considered the reference dataset and the other the incomplete dataset. However, different initial setups can be suitable as long as the initial incomplete and reference datasets are available.

Considering an original and complete dataset partitioned in two datasets—incomplete and reference databases—it is important to maintain a balanced representation of the original classes and feature values, especially in the reference dataset. The cluster validation step relies on the cluster distribution of the UMAP space. An imbalanced distribution may restrict the diversity of synthetic data generation, resulting in clusters that lack samples and thus do not allow for a reliable assignment of synthetic samples. Ensuring an even distribution across the reference dataset helps create a more robust synthetic dataset that accurately reflects the diversity of the original data.

2.3. Defining Value Ranges and Quantization

The next step defines the range (M) of possible synthetic values assigned to each feature. The range is obtained by analyzing the reference dataset, obtaining the minimum and maximum values and the step-size between different values. For categorical, logical, or ordinal variables, this is straightforward. However, continuous features require a quantization step to convert continuous values into discrete intervals suitable for controlled synthetic data generation. Quantization is a key element for continuous features to allow discrete, controlled synthetic values for comparison and validation.

2.4. Hyperparameter Tuning for Synthetic Data Validation

To optimize the methodology, specific hyperparameters must be fine-tuned for each experimental setup. The following hyperparameters are central to achieving effective, valid data generation:

- Density threshold for cluster area definition: cluster validation involves establishing a radius around each centroid within which synthetic samples are validated. The radius is adjusted until the sample density within it falls below a pre-set threshold and the percentage of points in the cluster within this radius meets an established threshold. The percentage threshold depends on the aim of the methodology; in this case, the interest is to define a cluster area that comprises a wide majority of the samples, and therefore the threshold is set to 90%. Synthetic samples falling within the validated cluster areas are then assigned to the corresponding cluster; those outside any cluster radius are not validated and thus discarded.

Since the density threshold determines the effective cluster area for synthetic data validation, it should be adjusted based on the reference data sample size and the shapes observed in the UMAP-reduced space. This adjustment ensures that the synthetic samples conform well to the reference dataset's inherent structure.

- ϵ for reliability score assignment: the reliability score quantifies the validity of each synthetic sample based on feature disparity. The feature disparity measure is the absolute difference between the synthetic value of a given feature and the mean value of the same feature among the nearest neighbors. However, depending on the attribute, the score is controlled by the hyperparameter ϵ, as shown in Table 1.

  The hyperparameter ϵ is determined by evaluating the mean feature disparity within the reference dataset. For each reference sample, the average feature value among the ten nearest neighbors is computed. The rounded version of the mean feature disparity is assigned to the feature ϵ.

2.5. Datasets

This subsection presents the different datasets used to verify the proposed algorithm. Within the health domain, three different topics have been chosen: prostate cancer, breast cancer, and cardiovascular data. By evaluating the methodology on different datasets, we aim to test the generalization of the algorithms across different domains, always focusing on health-related data.

2.5.1. Prostate Cancer

The dataset employed to evaluate the proposed algorithm is the publicly available Prostate Imaging and Cancer Analysis Information (PI-CAI) repository (PI-CAI database: pi-cai.grand-challenge.org, accessed on 15 December 2024), previously used in [24]. This dataset includes 1500 samples, each with 12 prostate cancer-related features. For the incomplete dataset setup, one attribute is systematically selected as missing and removed from the dataset to simulate incomplete data.

The primary target variable within this dataset is `case_csPCA`, which indicates the presence of clinically significant prostate cancer. A subset of features, including `patient_id`, `study_id`, and `mri_date`, are excluded from this analysis due to their lack of predictive value for `case_csPCA` and potential privacy concerns. Additionally, `histopath_type`, which identifies the lesion sampling method, is omitted as it does not directly inform the cancer diagnosis.

Key diagnostic features in PI-CAI include `lesion_GS`, which denotes the Gleason Score, a grading system used to evaluate prostate cancer aggressiveness by examining cell structure under a microscope. Similarly, `lesion_ISUP` consolidates the Gleason Score into five ISUP grades, aiding in clinical outcome prediction. The feature `case_ISUP` indicates the highest ISUP grade among sampled lesions, initially considered for analysis but later excluded due to a high correlation (0.87) with `case_csPCA`.

Further, `psa` represents the prostate-specific antigen level, and `psad` refers to the prostate-specific antigen density, which can also be calculated if missing by dividing `psa` by `prostate_volume` (`psad_computed`). Consequently, the primary predictive features for the target `case_csPCA` include `patient_age`, `prostate_volume`, and either `psa`, `psad`, or `psad_computed`.

2.5.2. Breast Cancer

This study includes two different breast cancer datasets. The first is the BC-MLR dataset, which was obtained from the UCI (University of Chicago, Chicago, IL, USA) Machine Learning Repository (BC-MLR dataset: https://archive.ics.uci.edu/dataset/14/breast+cancer, accessed on 15 December 2024). This dataset contains 286 instances with nine attributes that capture both patient-specific characteristics and tumor-related details.

For example, the dataset provides information on the patient's age at diagnosis (`age`) and menopausal status at that time (`menopause`). Tumour-specific characteristics include the maximum diameter of the tumor (`tumour size`) and the degree of malignancy (`deg-malig`).

In addition, the dataset covers cancer spread, noting the number of lymph nodes involved (`inv-nodes`) and whether the tumor has invaded the lymph node capsule (`node-caps`). Information about the side of the breast involved (`breast`) and the location of the tumor (`breast-quad`) is also included. The characteristic `irradiat` indicates whether the patient received radiotherapy. The target variable `Class` indicates whether the patient had a cancer recurrence after treatment.

The second dataset, known as BCNB, is the Early Breast Cancer Core-Needle Biopsy WSI dataset (BCNB dataset: https://paperswithcode.com/dataset/bcdalnmp, accessed on 15 December 2024). This dataset contains data from 1058 patients and includes comprehensive details of patient demographics and tumor characteristics. It records patient age at diagnosis (`age`) and tumor-related characteristics, such as `tumour size` and `tumour type`. Molecular markers including estrogen receptor (`ER`), progesterone receptor (`PR`), and human epidermal growth factor receptor 2 (`HER2`) expression are also recorded. In addition, `HER2 expression` provides immunohistochemistry test results to predict cancer response to treatment.

Additional features of the BCNB dataset include `histological grading`, which assesses how closely tumor cells resemble normal breast cells, and information on surgical interventions (`surgical`), specifically the type of lymphadenectomy performed. The dataset also includes the `Ki67 score`, which measures cancer cell proliferation, and the `molecular subtype`, a classification based on common tumor characteristics and hormone receptor status. Finally, lymph node involvement is detailed by features such as the number of metastatic lymph nodes and `ALN status`, which indicates whether the lymph nodes under the arm are involved.

Since we are aiming to work with both datasets for synthetic data generation, we focus on common features between the datasets: `age`, `tumor-size`, and `inv-nodes`.

2.5.3. Cardiovascular Disease

Two datasets are used in the field of cardiovascular disease. The first dataset, known as fiv-CardioDB, is a comprehensive collection of heart disease data from five different sources: Cleveland, Hungarian, Swiss, Long Beach VA, and the Statlog (Heart) Dataset. This dataset is from Liverpool John Moore's University (fiv-CardioDB dataset: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive, accessed on 15 December 2024) and contains 1190 instances with 11 features.

The second dataset, known as Ind-CardioDB, was developed by Doppala et al. at Lincoln University College (Ind-CardioDB dataset: https://data.mendeley.com/datasets/dzz48mvjht/1, accessed on 15 December 2024). This dataset consists of 1000 records with 12 characteristics collected from a hospital in India.

Both datasets have 11 features in common, which facilitates comparative analysis and usability. However, the Ind-CardioDB dataset includes an additional feature, the number of major vessels (`noofmajorvessels`), which is not present in fiv-CardioDB.

The common features include a variety of attributes, including general characteristics such as `age` and `sex`, as well as specific cardiovascular parameters. For example, `chest pain type` categorizes chest pain into four types: 1 for typical angina, 2 for atypical angina, 3 for non-anginal pain, and 4 for asymptomatic cases. Additional characteristics include `resting blood pressure` (measured in mm Hg), `serum cholesterol` (in mg/dL), and `fasting blood glucose` (considered true if > 120mg/dL).

Electrocardiographic data are collected by `resting electrocardiogram results` (`restingrelectro`), classified as 0 for normal, 1 for ST-T wave abnormality, and 2 for probable or definite left ventricular hypertrophy according to Estes criteria. The datasets also record `maximum heart rate achieved` (`maxheartrate`) and the binary attribute `exercise-induced angina` (`exerciseangia`).

Other attributes include `oldpeak`, which measures exercise-induced ST depression relative to rest, and `slope` of the peak exercise ST segment, categorized as 1—sloping, 2—flat, or 3—sloping. The target variable, `class`, indicates the presence or absence of heart disease.

2.6. Evaluation Metrics

The evaluation of synthetic data quality has evolved significantly over time. However, a universal method for benchmarking the performance of synthetic data generation has not yet been established. Several recent studies have aimed to review the diversity of metrics used for synthetic data validation, especially for tabular data, which often receive less attention compared with synthetic image data. For example, ref. [25] concentrates on the evaluation of synthetic tabular data, highlighting the need for specialized metrics in this domain. Importantly, the choice of validation metric depends on the intended application of the synthetic data.

Given that evaluation criteria vary based on the synthetic data's intended application, we select fidelity and utility metrics that align with the practical use of synthetic data in healthcare. Among the multiple synthetic data applications in the medical field, we aim to use synthetic data for machine learning model training when real data are insufficient due to privacy restrictions.

The fidelity of synthetic data is a measure of how well the synthetic data represent the statistical properties of real data. To assess fidelity, we use the cumulative distribution function (CDF), following the approach proposed by [26]. The CDF measures the probability that a variable will take a value less than or equal to a given point, allowing the comparison of the probability distributions of real and synthetic datasets. The Kolmogorov–Smirnov (K-S) test, as in [27–29], is used to measure this similarity. The K–S test calculates a statistic based on the maximum absolute difference between the CDFs of the real and synthetic datasets. This maximum difference serves as the fidelity score, where a lower value indicates that the synthetic data distribution closely aligns with that of the real data, thereby demonstrating high fidelity.

On the other side, a significant element of synthetic data quality is its utility, which is commonly assessed through the examination of the performance of machine learning models trained on synthetic data. This method of evaluation was initially proposed by Esteban et al. [30] under the framework "Train on Synthetic, Test on Real" (TSTR). In this approach, machine learning models are trained separately on real and synthetic data, and then tested on a common holdout dataset that has not been seen by either the models or the generative algorithm. This comparison of model performance between real and synthetic data has become a prevalent approach in the literature, as evidenced by [26,31–33].

An alternative approach to evaluating data utility involves assessing the performance of models trained with augmented data (a combination of real and synthetic samples). Data augmentation increases the diversity and quantity of training data by incorporating synthetic samples [34]. This methodology has proven especially valuable in the health domain, addressing critical challenges such as class imbalance by generating additional samples for underrepresented classes, improving model generalization by exposing models to a broader range of data and reducing overfitting, and mitigating bias by creating more representative datasets that include minority groups [35]. Class imbalance, in particular, has been shown to result in unfair decisions for minority classes, as highlighted in [36].

Consequently, a key utility metric involves comparing the performance of models trained exclusively on real data against those trained on augmented datasets. This approach has been successfully demonstrated in previous studies, such as [37], where augmented datasets were used to enhance model performance and fairness.

A variety of machine learning models can be used for this evaluation. In this paper we focus on random forests, logistic regression, and support vector machine (SVM) networks. These models are trained with real, synthetic, and augmented data and then tested on real data.

The effectiveness of each ML model is measured using three common evaluation metrics: accuracy, F1-score, and area under the curve (AUC). Accuracy represents the percentage of correctly classified samples out of the total samples, providing a straightforward measure of overall model performance. The F1-score balances precision and recall, especially important when dealing with imbalanced classes often present in medical datasets, and AUC measures the ability of the model to distinguish between classes, with higher AUC values indicating better class separability.

To evaluate the quality of our UMAP-based synthetic data generation algorithm, the obtained fidelity and utility metrics are compared with the analog measures for three widely used state-of-the-art methods: CTGAN, CopulaGAN, and TVAE. These methods are commonly applied in healthcare data generation [38–40] and provide effective benchmarks for assessing the performance of our algorithm.

CTGAN [41], or Conditional Tabular GAN, uses a Generative Adversarial Network framework designed specifically for synthetic tabular data generation. It addresses several challenges in GAN-based data generation, such as mixed data types, complex multi-modal distributions, and class imbalance in categorical variables. CTGAN incorporates mode-specific normalization to capture patterns without mode collapse and employs a conditional generator to better represent minority classes in categorical variables, allowing it to closely reproduce real-world data distributions. This architecture includes separate pre-processing steps for continuous and discrete columns and uses fully connected networks in both the generator and discriminator, enabling CTGAN to handle intricate relationships among data variables.

In contrast, CopulaGAN [42] combines GANs with copula functions [43], which are mathematical models used to capture dependencies between variables independently of their marginal distributions. By leveraging Gaussian copulas, CopulaGAN transforms data distributions into Gaussian-like forms, simplifying the learning process for complex tabular data. This approach allows CopulaGAN to better represent non-linear dependencies in datasets, making it particularly effective for applications requiring the preservation of inter-variable relationships.

The TVAE, also referred to as a tabular variational autoencoder, firstly proposed in [41], extends the Variational Autoencoder (VAE) framework for tabular data. Unlike traditional autoencoders, which map input data to deterministic latent representations, VAEs use a probabilistic latent space, representing data as distributions rather than fixed points [15]. This enables sampling from the latent space to generate new data points consistent with the input distribution. The TVAE adapts this approach to handle the unique challenges of tabular data, such as mixed data types and feature dependencies.

2.7. Ethical Considerations for Synthetic Data in Sensitive Domains

Synthetic data offer significant opportunities to address biases in healthcare datasets and alleviate privacy concerns, as previously discussed. However, their use raises important ethical challenges that require careful consideration, including issues related to fairness, privacy, and potential misuse [44]. These challenges highlight the need for a nuanced understanding of the limitations and implications of synthetic data in sensitive domains such as healthcare [45,46].

Fairness remains a central concern when using synthetic data. Despite their potential, synthetic data are not inherently unbiased. The generation process relies on real datasets, which often contain patterns and biases embedded in the original data [47]. Consequently, any bias present in the source data is likely to be replicated, or even amplified, in the synthetic dataset. Efforts to mitigate such biases have led to the development of fairness-aware techniques, which could theoretically be applied to both real and synthetic data to address these disparities [48].

Privacy considerations also require significant attention. Although synthetic data are often perceived as a privacy-preserving solution, this perception can be misleading. As noted by the Royal Society and The Alan Turing Institute, “Synthetic data are not automatically private" [49]. While synthetic datasets may obscure direct identifiers, they are not entirely exempt from privacy risks and must still be handled with caution. The illusion of privacy can lead to complacency, but safeguards must be implemented to prevent potential privacy breaches.

Transparency and explainability are equally crucial when applying synthetic data generation methods, particularly in healthcare. Clearly documenting the entire process—from algorithm selection and parameter tuning to assumptions and challenges encountered—fosters trust and accountability. This documentation should provide stakeholders with detailed insights into how synthetic datasets were created, including any trade-offs made during the process. Transparency ensures that end-users of synthetic data can understand their limitations and make informed decisions about their use.

Best practices for responsible synthetic data usage include regular bias assessments to identify and mitigate any disparities [47]. Transparent reporting is also essential, requiring clear labeling of synthetic datasets and explanations of the generation process to ensure users understand the data's provenance and limitations [50]. Additionally, the balance between data utility and privacy protection must be carefully evaluated to maximize the effectiveness and ethical alignment of synthetic data applications.

It is also crucial to recognize that synthetic data cannot fully replace real-world datasets [49]. By design, synthetic data introduce some level of distortion to ensure privacy or mitigate biases. These distortions, while necessary, may affect the validity of downstream analyses and model training. As such, synthetic data should primarily serve as a complementary resource to accelerate the research pipeline. Ultimately, models or tools intended for real-world deployment must be validated and fine-tuned using real data to ensure their reliability and robustness.

3. Results

This section presents the outcomes of applying the proposed algorithm to generate fully synthetic datasets for three medical domains: prostate cancer, breast cancer, and cardiovascular disease.

3.1. UMAP Transformation and Feature Distribution

The first phase of the methodology relies on transforming the input data into a lower-dimensional space using UMAP. This transformation provides a two-dimensional representation that facilitates the visualization of feature distributions within the datasets. Such visualizations are critical for validating the quality of the synthetic data generated. However, UMAP is known to be sensitive to its initialization and hyperparameters, which can significantly affect the resulting cluster structure and data distribution in low-dimensional space. To address this sensitivity, we performed a detailed sensitivity analysis using prostate, breast, and cardiovascular datasets.

UMAP's key parameters include `n_neighbors`, `min_dist`, and `random_state`. The `n_neighbors` parameter controls the balance between local and global data structure, determining the size of the neighborhood UMAP considers when learning the data manifold. The `min_dist` parameter controls how closely UMAP packs points in the low-dimensional representation, effectively setting a lower bound on the distance between points. Finally, the `random_state` parameter affects the randomness of the initialization, allowing reproducibility when set to a fixed value.

To assess the impact of these parameters, we performed 50 UMAP transformations for each dataset, varying the values of `n_neighbors`, `min_dist`, and `random_state`. Specifically:

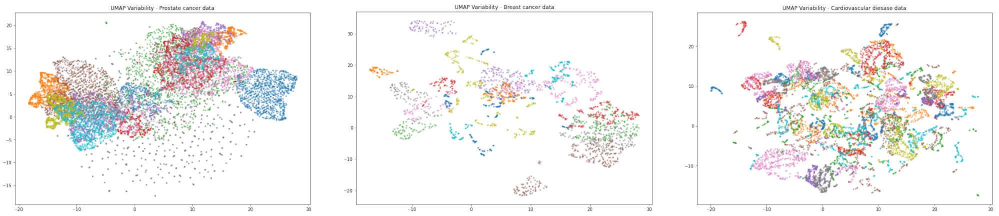
- `n_neighbors` parameter was randomly assigned integer values between 2 and 50.
- `min_dist` parameter was sampled from a uniform distribution between 0.0 and 0.9.
- `random_state` was randomly chosen as an integer between 0 and 10,000.

Each combination of these parameters resulted in a unique UMAP transformation, producing different data distributions in two-dimensional space. To quantify the variability across these transformations, we computed the cluster centroids for each class in each dataset and calculated the variance in their coordinates across the 50 transformations. The results, presented in Table 2, show considerable variability in the centroid positions, with the smallest variance being 18.68. This significant variation highlights the sensitivity of UMAP to initialization and hyperparameters, which can lead to inconsistent cluster structures depending on the settings chosen.

Table 2. Variance on the cluster centroid coordinates for 50 different UMAP initializations.

	Cluster 0.0		**Cluster 1.0**	
	x coord.	y coord.	x coord.	y coord.
Prostate cancer data	173.99	18.68	45.99	50.45
Breast cancer data	99.06	53.22	87.81	101.39
Cardiovascular diseases	60.33	146.43	21.15	19.36

To further illustrate this variability, Figure 4 displays 10 random UMAP transformations for each dataset. These visualizations show that, while some transformations maintain clear cluster separation, others fail to preserve the underlying structure of the data. For example, in several transformations, the two clusters remain clearly separated, but their spatial arrangement varies significantly. In other cases, suboptimal initializations result in poor cluster definition, compromising the interpretability of the representation.



(**a**) Prostate cancer data transformation (PI-CAI)

(**b**) Breast cancer data transformation (BC-MLR)

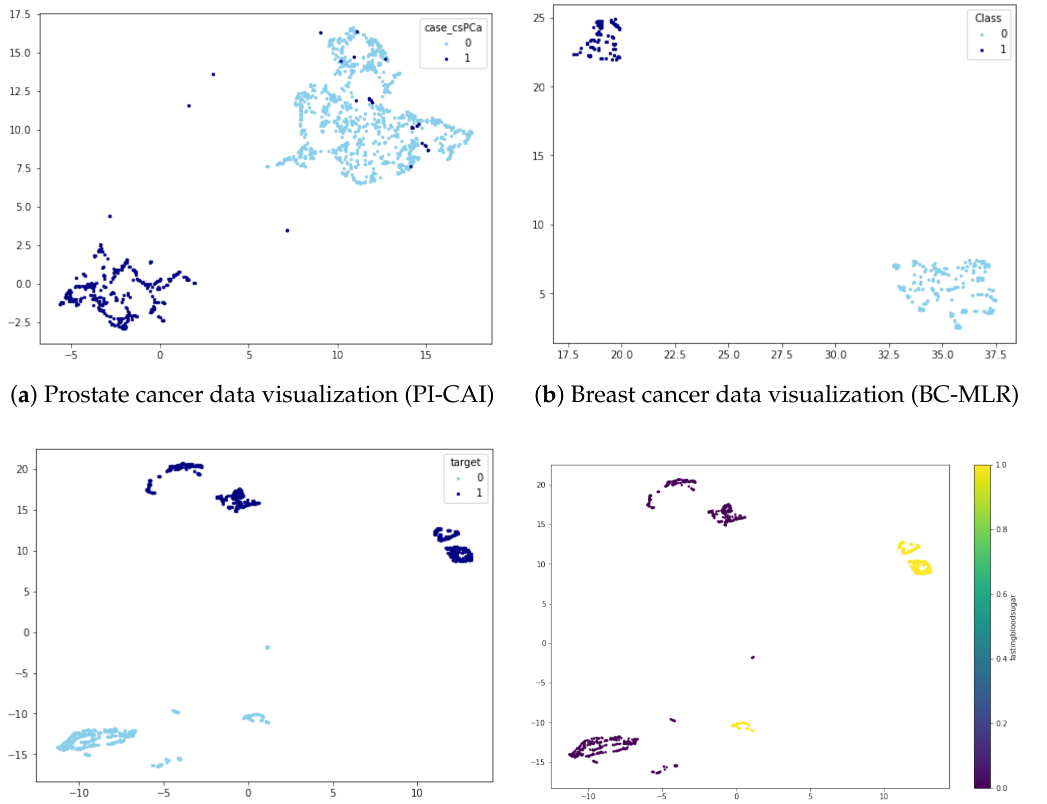
(**c**) Cardiovascular diseases data transformation (Ind-CardioDB)

Figure 4. UMAP transformation for 10 different initializations. It is highlighted using different colors how samples are grouped according to the target variable.

This sensitivity analysis highlights the importance of carefully selecting UMAP parameters to ensure meaningful and reproducible results, especially when used for synthetic data validation and clustering tasks. Figure 5 illustrates the UMAP-transformed representations of three different datasets for the appropriate initialization. For these datasets, the chosen parameters are `n_neighbors`= 15 and `min_dist`= 0.1. The `random_state` values were set to 23 for the prostate cancer data, 80 for the breast cancer data, and 76 for the cardiovascular disease data to ensure reproducibility.

Specifically, Figure 5a displays the UMAP transformation of the PI-CAI prostate cancer dataset, highlighting how samples are grouped according to the target variable, (`case_csPCA`). Similarly, Figure 5b presents the UMAP visualization of the BC-MLR breast cancer dataset, showing distinct clusters that correspond to the target feature, `Class`. Finally, Figure 5c depicts the UMAP-transformed Ind-CardioDB cardiovascular dataset, where samples are also distinctly grouped based on the target variable.

In the case of the cardiovascular data (Figure 5c), an additional level of differentiation is observed within the primary clusters. Further analysis reveals that this secondary separation corresponds to the feature `fastingbloodsugar`, as shown in Figure 5d. This highlights UMAP's ability to capture subtle structures within the data.



(**a**) Prostate cancer data visualization (PI-CAI)

(**b**) Breast cancer data visualization (BC-MLR)

(**c**) Cardiovascular diseases data visualization (Ind-CardioDB)

(**d**) `fastingbloodsugar` value distribution in the Cardiovascular dataset (Ind-CardioDB)

Figure 5. Reference datasets visualization of the UMAP transformation.

3.2. Dataset Partitioning and Setup

For the prostate cancer dataset, the primary goal is to generate a fully synthetic dataset based on the original PI-CAI data. The algorithm is applied by splitting the dataset into a reference dataset and an incomplete dataset. The features of interest for this dataset include `patient_age`, `prostate_volume`, and `psa` (chosen over alternatives such as `psad` or `psad_computed`). Additionally, the target variable `case_csPCA` is included.

The PI-CAI dataset comprises 1500 samples with 425 samples representing `case_csPCA` = 1 and 1075 representing `case_csPCA` = 0. While the dataset is imbalanced, the `case_csPCA` = 1 class is sufficiently represented within the UMAP-transformed space, as seen in Figure 5a. The data are divided into a reference dataset (500 samples) and an incomplete dataset (1000 samples), maintaining a similar target variable proportion, with approximately 28.3–28.4% positive samples across both datasets.

Regarding the breast cancer and cardiovascular disease data, two datasets are available, and therefore the division of data is not necessary. Instead, each dataset can be considered as either reference or incomplete. In the case of breast cancer, the BC-MLR dataset, which is considered public and privacy protected, is referred to as the reference dataset, while the BCNB dataset is considered incomplete. The initial focus is on generating three common features shared by both datasets: `age`, `tumor-size`, and `inv-nodes`. However, the iterative nature of the algorithm allows for the generation of additional features present in the reference dataset. Based on correlation analysis with the target variable `class`, three additional attributes are prioritized: `node-caps` (correlation: 0.24), `deg-malig` (0.3), and `irradiat` (0.19). The final goal is to generate these six attributes along with the target variable.

With respect to the cardiovascular disease datasets, the Ind-CardioDB dataset is referred to as the reference dataset, while fiv-CardioDB serves as the incomplete dataset.

Both datasets share most attributes, which facilitates comparative analysis. A correlation analysis with the target variable `target` was performed for both datasets. Interestingly, significant differences were observed in the correlation with the output variable of identical attributes between the two datasets. Finally, four attributes were selected for the fully synthetic dataset based on their average correlation in both datasets: `chestpain` (0.51), `restingBP` (0.3), `fastingbloodsugar` (0.28), and `maxheartrate` (0.32).

3.3. Defining Value Ranges and Quantization

The following value ranges were identified for each attribute of the prostate cancer dataset:

- `patient_age`: categorical variable. Integer values from 35.0 to 92.0.
- `psa`: continuous variable. Decimal values from 0.10 to 224.00.
- `prostate_volume`: categorical variable. Integer values from 4.0 to 308.0. A few values contain decimals, but we assume all integers.
- `case_csPCa`: logical variable. Possible values 1.0 and 0.0.

Quantization is applied to the continuous variable `psa` from the PI-CAI dataset, to transform its values into discrete categories. Figure 6 illustrates the original PSA values. Each PSA value is rounded to the nearest integer or 0.5 step, ensuring a discrete representation. Figure 7a shows the correspondence between quantized and original continuous values, with Figure 7b providing a focused view for values under 20. Values exceeding 100 (0.4% of the samples) are treated as outliers and removed from the quantization range. This consideration is performed given the very low percentage of samples and considering the nature of this specific dataset. In this case, a psa above 10 ng/mL is already considered abnormal. However, these types of considerations should be carefully performed, as outliers represent an important font of information, specially in healthcare data. After the quantization, the M range for the `psa` variable is from 0 to 100 with a step of 0.5 (instead of 1, as in the integer ranges).

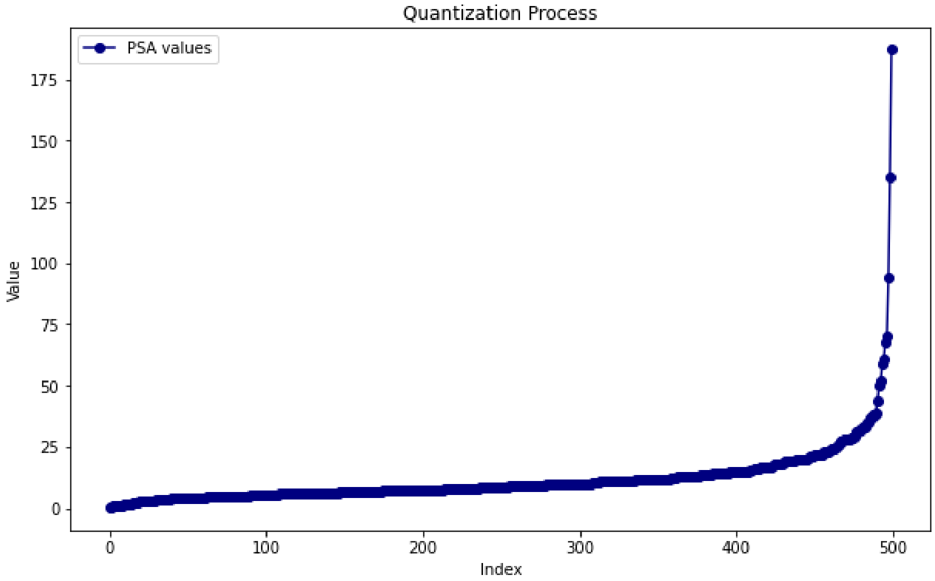


Figure 6. PSA values.

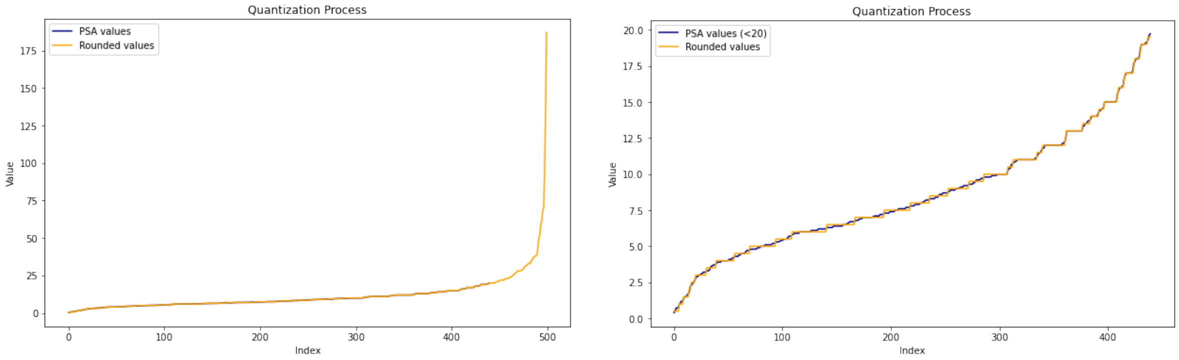
For the breast cancer data, the values needed to be encoded prior to defining the ranges. After encoding, the value ranges for each variable were obtained:

- `age`: categorical variable. Integer values from 0.0 to 5.0.
- `tumor-size`: categorical variable. Integer values from 0.0 to 10.0.
- `inv-nodes`: categorical variable. Integer values from 0.0 to 6.0.
- `node-caps`: logical variable. Possible values 0.0 and 1.0.
- `deg-malig`: categorical variable. Integer values from 1.0 to 3.0.

- `irradiat`: logical variable. Possible values 0.0 and 1.0.
- `Class`: logical variable. Possible values 0.0 and 1.0.

The ranges of values for the cardiovascular dataset are as follows:

- `chestpain`: categorical variable. Integer values from 1.0 to 4.0.
- `restingBP`: categorical variable. Integer values from 94.0 to 200.0.
- `fastingbloodsugar`: logical variable. Integer values from 0.0 to 1.0.
- `maxheartrate`: categorical variable. Possible values 71.0 and 202.0.
- `target`: logical variable. Possible values 0.0 and 1.0.



(**a**) Original values and quantization process (**b**) Quantization focus (values < 20)

Figure 7. Quantization process applied to the PSA values.

In these cases, there is no need to apply quantization because none of the variables from BC-MLR and Ind-CardioDB are continuous.

3.4. Hyperparameter Tuning for Synthetic Data Validation

Hyperparameters, such as the density threshold for defining cluster areas and the ϵ for assigning reliability scores, were tuned based on UMAP-transformed data distributions.

The density threshold plays a key role in determining the cluster validation process. Cluster validation depends on the cluster area, which is directly affected by the selected density threshold. A lower density threshold expands the cluster area, allowing more samples to achieve cluster validation. Conversely, as the density threshold increases, the cluster area shrinks, reducing the number of validated samples.

In order to determine the optimal density threshold, the percentage of cluster validated samples against the different density thresholds is considered, as shown in Figure 8. It can be observed how a high percentage of validated samples is achieved up to the elbow point, after which it decreases. Beyond this point, the cluster area becomes too small, resulting in fewer samples achieving validation. This trend highlights the importance of selecting a threshold that maximizes the inclusion of meaningful samples without compromising the accuracy of cluster validation. With this aim, the elbow point for each dataset has been computed, resulting in a chosen density threshold equal to 4 for prostate and breast cancer data and equal to 5 for cardiovascular disease data.

As an example, Figure 9 illustrates cluster areas defined by different density thresholds in the breast cancer data. For thresholds above the selected value, the cluster area does not include all samples within the cluster. Conversely, lower density thresholds produce larger areas, resulting in validation of samples that deviate significantly from the cluster in UMAP space.

The ϵ hyperparameter for reliability scores is defined based on the average feature disparity among reference points, and serves as a threshold for assessing the reliability of the generated samples. A generated sample with a feature disparity close to or below ϵ will receive a high reliability score, indicating that its feature values are consistent with those of reference points.

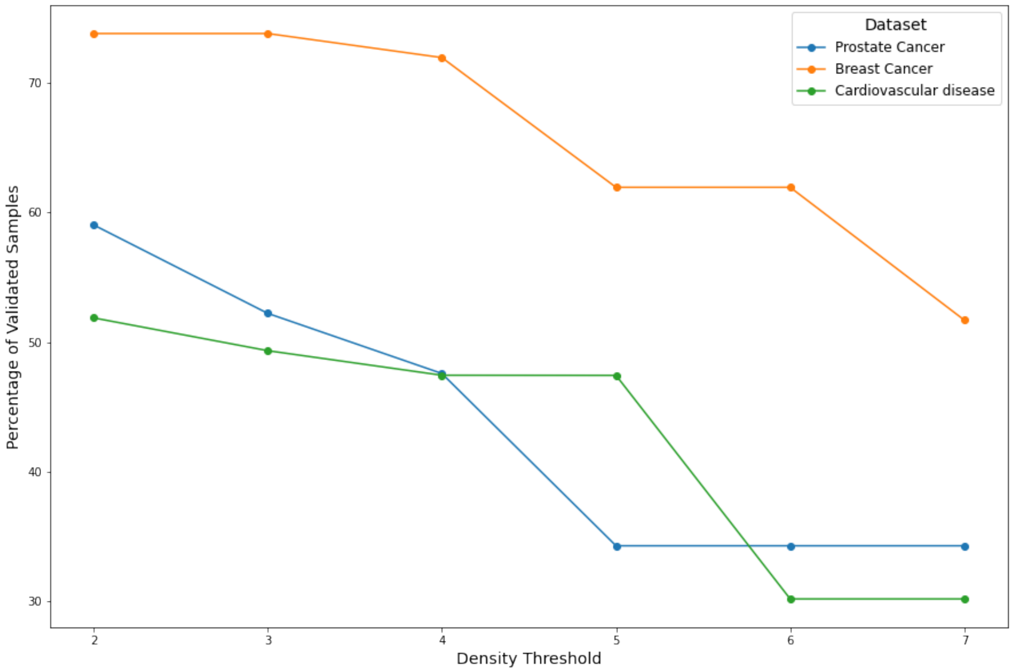


Figure 8. Percentage of cluster-validated samples for different density thresholds.

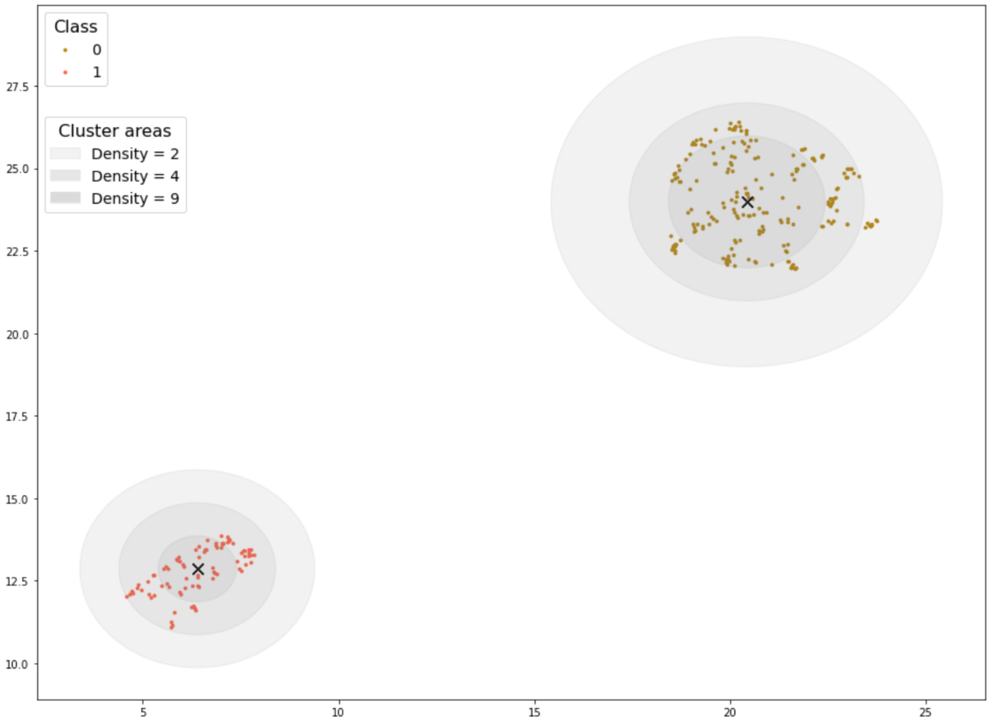


Figure 9. UMAP visualization of breast cancer reference data (BC-MLR) and cluster areas for different density thresholds.

To calculate ϵ, the feature disparity is computed by evaluating the mean difference between each reference point's feature value and the average feature value of its 10 nearest neighbors in the UMAP-transformed space. This mean feature disparity serves as the foundation for ϵ adjustment, ensuring that the threshold is data-specific and reflective of the distribution of the reference dataset.

Table 3 presents the computed mean feature disparity values for each feature in the prostate cancer dataset as an example of this process. It can be observed how ϵ is the closest integer to the mean feature disparity values. An analog procedure was applied to the other datasets analyzed in this study to ensure consistency and accuracy in defining the ϵ values for each case.

Table 3. Mean feature disparcity for prostate cancer reference data (PI-CAI).

Feature	Mean f_{disp}	ε
`patient_age`	1.18	1
`PSA`	3.77	4
`prostate volume`	5.67	6

3.5. Iterative Synthetic Data Generation Results

With the pre-processed data and optimized hyperparameters, the iterative application of the proposed algorithm successfully generated three fully synthetic datasets, each tailored to the characteristics of the input data. The process involved generating synthetic values for different attributes in successive iterations, culminating in the generation of the target variable based on the UMAP-transformed distribution of the generated samples.

For the prostate cancer dataset (PI-CAI), the algorithm started with an incomplete dataset of 1000 entries. After iteratively generating synthetic attributes and applying reliability score filtering and cluster validation, the final fully synthetic dataset consisted of 79,033 samples.

In the case of breast cancer data, the BCNB dataset served as the incomplete dataset, containing 1058 samples with three attributes. In this case, since part of the input dataset did not contain all the attributes we aimed to generate, the partially synthetic data generation method [24] was applied to generate the missing attributes. This resulted in a partially synthetic dataset of 9427 samples with six attributes and one target variable. Using this partially synthetic dataset as the incomplete dataset, the proposed iterative algorithm was applied to generate a complete fully synthetic dataset of 17,225 samples.

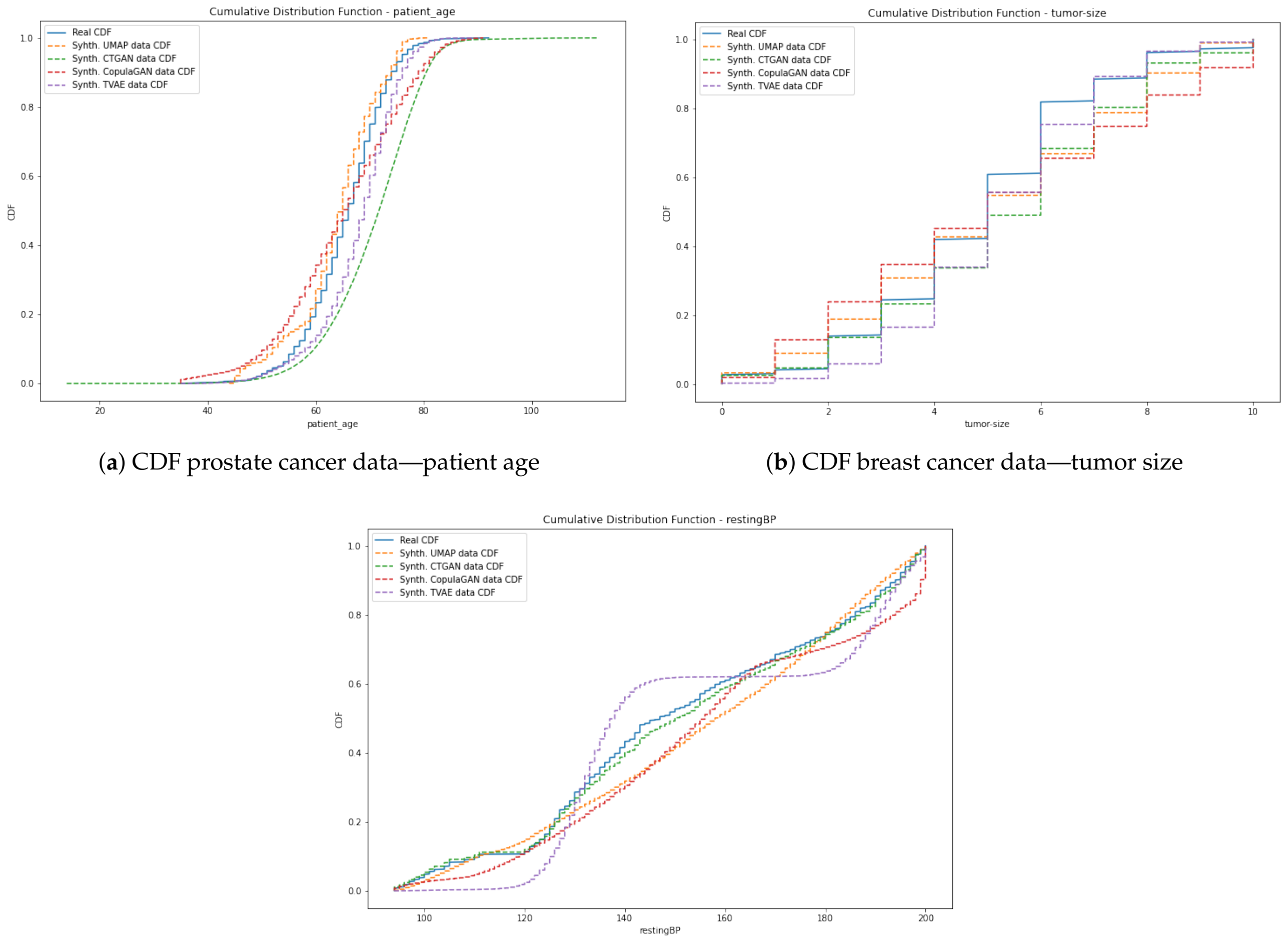
For the cardiovascular disease dataset, the initial dataset (fiv-CardioDB) contained 1190 samples. By iteratively generating synthetic values for all attributes, the algorithm produced a fully synthetic dataset of 50,476 samples.

3.6. Evaluation Metrics

After generating the synthetic datasets, a quality assessment is conducted to evaluate the effectiveness of the UMAP-based approach. To establish a performance benchmark, we compare our results with the evaluation metrics obtained from synthetic datasets generated using CTGAN, CopulaGAN, and TVAE. These synthetic datasets have been created from the reference dataset in each case, with an output size matching the scale of the UMAP-based synthetic data (79,000 for prostate cancer, 17,000 for beast cancer, and 50,000 for cardiovascular diseases).

The evaluation metrics encompass two key dimensions: fidelity and utility. Fidelity is assessed by comparing the cumulative distribution function (CDF) of real and synthetic data, measuring how well the synthetic data represent the distribution of the original dataset. Utility is evaluated based on the synthetic data's contribution to machine learning (ML) model performance, reflecting its practical applicability for data-driven healthcare tasks.

The fidelity assessment of the synthetic data generation through CDF is performed both visually, as shown in Figure 10, and quantitatively, using the Kolmogorov–Smirnov statistic, which measures the maximum absolute difference between the synthetic and real data distributions. This K-S statistic, presented in Table 4, provides a robust metric for evaluating distributional similarity for each attribute.



(**a**) CDF prostate cancer data—patient age

(**b**) CDF breast cancer data—tumor size

(**c**) CDF cardiovascular diseases data—resting BP

Figure 10. Cumulative distribution function per attribute for real data, UMAP, CTGAN, CopulaGAN, and TVAE synthetic data.

Table 4. K-S statistic for CDFs. Lower K-S statistic for each attribute in bold.

	Attribute	UMAP	CTGAN	CopulaGAN	TVAE
Prostate Cancer	`patient_age`	0.1587	0.3299	**0.1554**	0.1673
	`psa`	0.4103	0.4695	0.3583	**0.1980**
	`prostate_volume`	0.3639	0.3557	0.1690	**0.1327**
Breast Cancer	`age`	**0.3262**	0.3990	0.3289	0.5524
	`tumor-size`	0.1793	**0.1344**	0.2042	0.1410
	`inv-nodes`	**0.6010**	0.6942	0.6254	0.9477
	`node-caps`	**0.4254**	0.7501	0.6662	0.8906
	`tumor-size`	0.1793	**0.1344**	0.2042	0.1410
	`deg-malig`	**0.3756**	0.4090	0.3876	0.5921
	`irradiat`	**0.3931**	0.7118	0.6330	0.9448
Cardiovascular Diseases	`chestpain`	**0.3505**	0.4443	0.3834	0.4865
	`restingBP`	0.1351	**0.0382**	0.1419	0.1478
	`fastingbloodsugar`	**0.5749**	0.6689	0.6730	0.8934
	`maxheartrate`	0.0568	**0.0404**	0.1919	0.3230

Figure 10 shows the CDFs for various attributes across the datasets. For the prostate cancer data, the `patient_age` attribute is shown in Figure 10a. In the case of the breast cancer datasets, the CDF of the `tumor-size` attribute is illustrated in Figure 10b. In both plots, none of the synthetic datasets perfectly replicate the real data distribution at every point. However, the overall behavior of the four synthetic datasets is similar, with comparable deviations from the real CDF. This indicates that, while minor discrepancies exist, the general distributional properties are preserved across all methods.

For the cardiovascular disease datasets, the CDF of `restingBP` is depicted in Figure 10c. In this case, the synthetic datasets generated by UMAP and CopulaGAN closely follow the real data distribution, demonstrating high fidelity. Conversely, the synthetic datasets produced by CTGAN and TVAE show slight deviations from the real cumulative function, characterized by a subtle sinusoidal pattern.

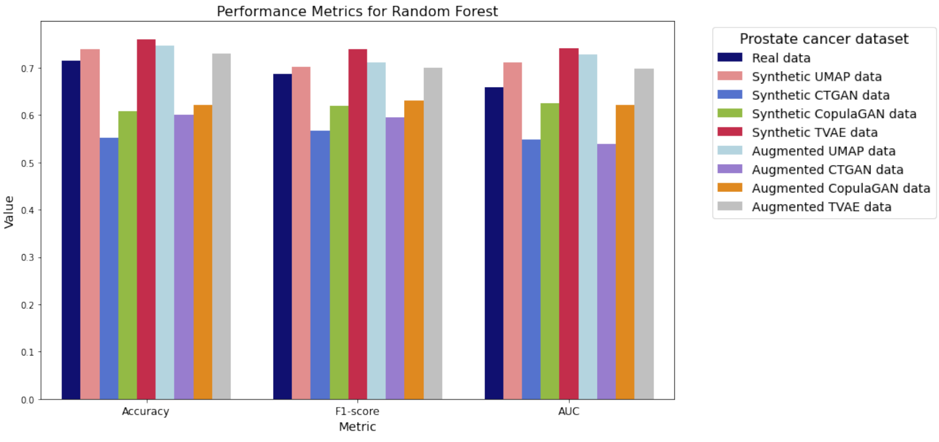
The quantitative results of the fidelity assessment, as measured by the Kolmogorov–Smirnov (K-S) statistic, are presented in Table 4, with the lowest values highlighted in bold. For the prostate cancer dataset, the synthetic datasets generated by CopulaGAN and TVAE demonstrate the smallest differences between the real and synthetic distributions, indicating superior fidelity for this case. Conversely, for the breast cancer data, the UMAP-generated and CTGAN synthetic datasets yield better fidelity results. Notably, the UMAP-based synthetic data achieve the lowest K-S statistic for most attributes in the breast cancer dataset, emphasizing their ability to closely replicate the real data distributions in this scenario.

The utility assessment evaluates the performance of three machine learning models—random forest classifier, logistic regression, and support vector machine—in predicting the target variable for each dataset (`case_csPCa`, `Class`, or `target`). Each model is trained on three types of datasets: real, synthetic, and augmented (a combination of real and synthetic data). For the synthetic and augmented cases, the training incorporates datasets generated using UMAP-based methods, CTGAN, CopulaGAN, and TVAE, resulting in nine different training datasets that will lead to different results.

The results of the random forest classifier are presented in Figure 11. For the prostate cancer data (Figure 11a), the UMAP-generated and TVAE synthetic and augmented datasets show superior performance compared with training on real data alone or training with CTGAN and CopulaGAN synthetic and augmented datasets.

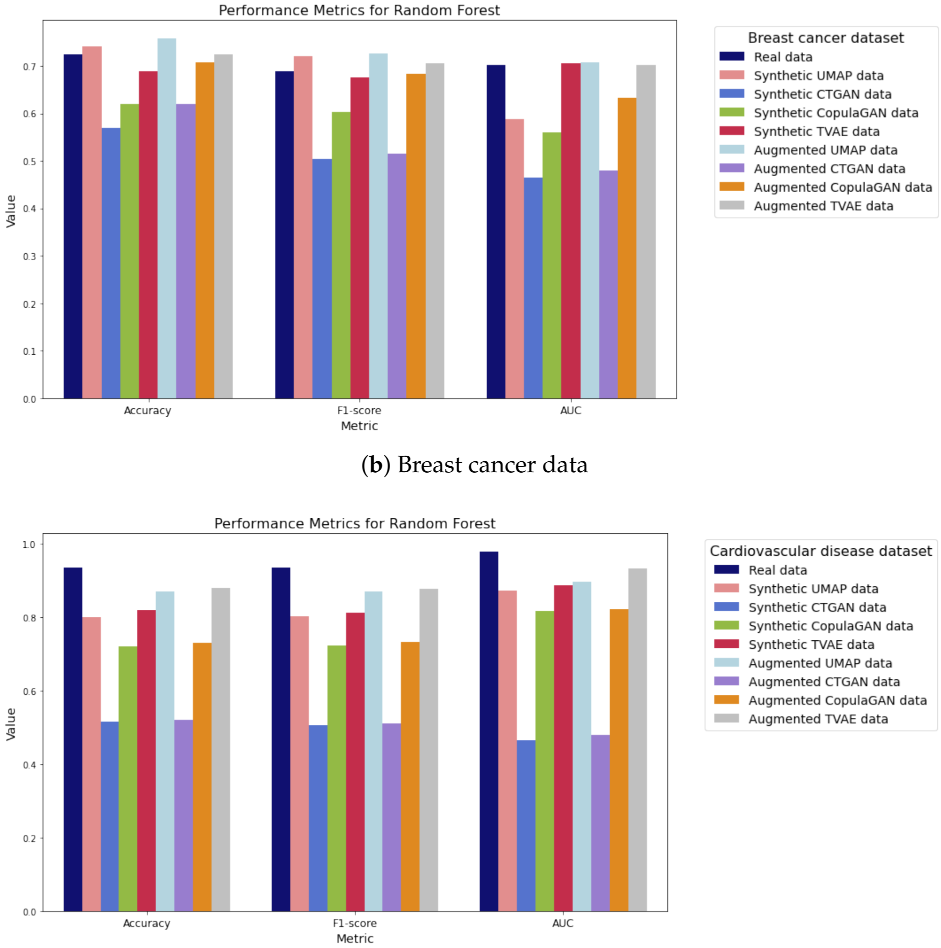
For the breast cancer data (Figure 11b), the performance enhancement provided by TVAE is less pronounced. However, UMAP synthetic and augmented datasets exhibit better results than real data across all metrics, except for AUC. In this specific case, only the augmented UMAP and TVAE datasets match the AUC performance achieved by real data.

In the case of the cardiovascular data (Figure 11c), none of the synthetic or augmented datasets surpass the performance of the real dataset. Additionally, the CTGAN-generated dataset demonstrates particularly poor performance metrics for this scenario.



(**a**) Prostate cancer data

Figure 11. *Cont.*



(**b**) Breast cancer data

(**c**) Cardiovascular diseases data

Figure 11. ML performance metrics (accuracy, F1-score, and AUC) for random forest model trained with real, synthetic, and augmented data.

The performance metrics for the logistic regression models, as displayed in Figure 12, reveal distinct patterns across the datasets. In the prostate cancer case (Figure 12a), while UMAP- and TVAE-generated datasets outperform CTGAN and CopulaGAN, they do not consistently exceed the performance of models trained on real data.

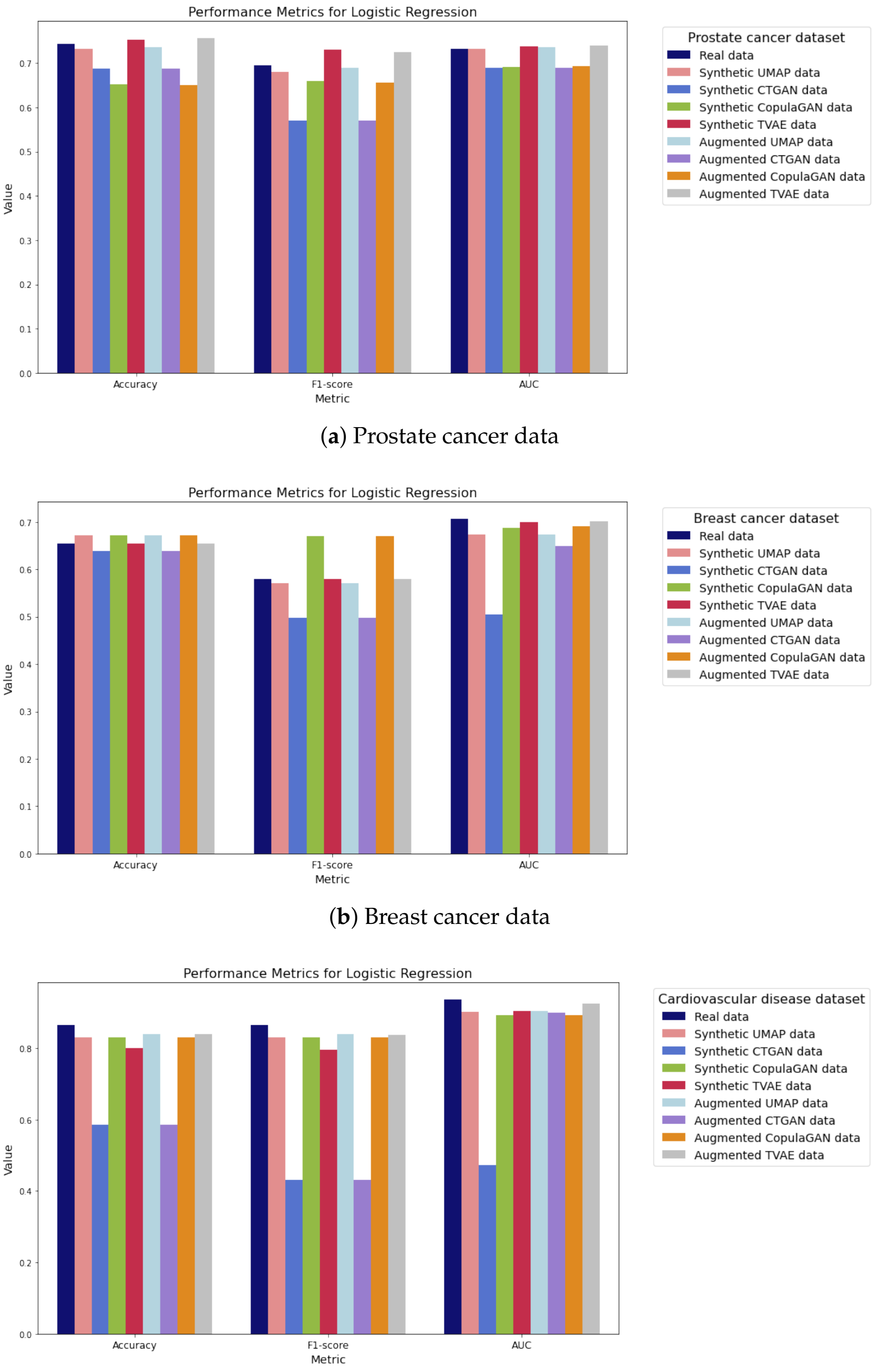
For breast cancer data (Figure 12b), CopulaGAN-generated datasets demonstrate improved performance compared with previous observations, particularly in the F1-score metric. Additionally, both CopulaGAN- and UMAP-generated datasets slightly outperform the real dataset in accuracy.

In the cardiovascular diseases scenario (Figure 12c), similar to the random forest results (Figure 11c), none of the synthetic datasets outperform the model trained with real data.

Support vector machine (SVM) results, presented in Figure 13, provide further insights. For prostate cancer data (Figure 13a), synthetic and augmented datasets perform similarly, slightly exceeding the performance of real data across most metrics. Notably, the F1-score demonstrates a significant improvement with synthetic datasets.

In the breast cancer case (Figure 13b), no synthetic dataset clearly outperforms the real data model across all metrics. However, subtle improvements in the F1-score are observed for some synthetic and augmented datasets, and augmented TVAE data show a marked enhancement in AUC performance.

Contrasting with the previous models, the cardiovascular diseases scenario (Figure 13c) showcases superior performance for synthetic and augmented datasets compared with real data. Among these, the UMAP-generated data stand out, delivering excellent results and outperforming all other datasets across the evaluated metrics.

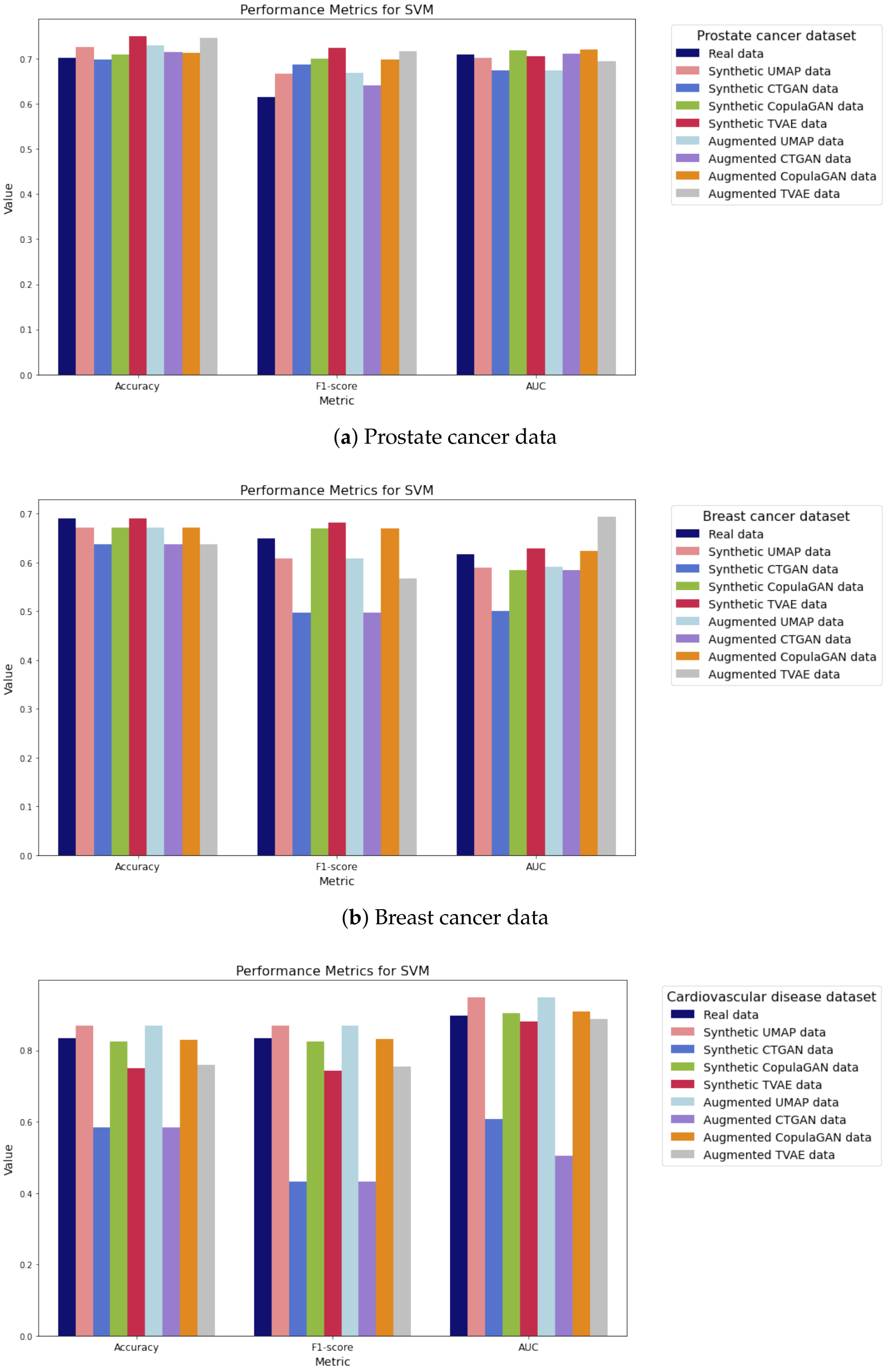


(**a**) Prostate cancer data

(**b**) Breast cancer data

(**c**) Cardiovascular diseases data

Figure 12. ML performance metrics (accuracy, F1-score, and AUC) for logistic regression model trained with real, synthetic, and augmented data.



(**a**) Prostate cancer data

(**b**) Breast cancer data

(**c**) Cardiovascular diseases data

Figure 13. ML performance metrics (accuracy, F1-score, and AUC) for support vector machine model trained with real, synthetic, and augmented data.

4. Discussion

In this study, we have successfully extended a previously developed partially synthetic data generation methodology to create fully synthetic datasets across three different domains. Building on the UMAP-based generation algorithm, we achieved not only complete synthetic datasets but also significant improvements in data quality, as validated against state-of-the-art methods such as CTGAN, CopulaGAN, and TVAE. Our findings demonstrate that the UMAP-based method outperforms these benchmarks in terms of fidelity, particularly for breast cancer and cardiovascular disease data. Moreover, the synthetic and augmented datasets generated using the UMAP-based approach exhibit high utility, as evidenced by enhanced machine learning model performance across different models.

In terms of fidelity, the UMAP-based method outperformed other state-of-the-art techniques in most features across the datasets, as assessed by the Kolmogorov–Smirnov (K-S) test for cumulative distribution function evaluation. While the method consistently achieved high fidelity across all domains except for prostate cancer data, the differences in K-S statistics between UMAP and other methods were generally slight for most attributes. Nonetheless, the results underline the competitive performance of the UMAP-based methodology in ensuring fidelity while generating synthetic data.

The UMAP-based synthetic and augmented datasets also demonstrated promising results in terms of utility. Across most machine learning models and dataset domains, the UMAP-generated data exhibited enhanced predictive capabilities compared with models trained exclusively on real data. Of particular note is the performance of the SVM model with cardiovascular data, where the UMAP-based synthetic datasets outperformed all other datasets, including real, synthetic, and augmented ones. In other cases, TVAE datasets also showed strong performance, often comparable to the UMAP-based data. Nevertheless, it is important to note that, while model performance improved with synthetic data, the gains remain moderate, suggesting that further tuning or larger reference datasets could further optimize model outcomes.

As highlighted in the results analysis, augmented datasets often yield better utility outcomes compared with real or synthetic datasets alone. Data augmentation serves as a powerful tool for increasing both the diversity and volume of training data, thereby enhancing model performance. An intriguing direction for future research would be to systematically analyze the impact of different augmentation ratios, aiming to identify the optimal balance between real and synthetic data for various machine learning tasks.

An interesting finding is the low correlation observed between fidelity and utility metrics. Higher fidelity, as measured by the Kolmogorov–Smirnov statistic, does not necessarily correspond to better utility metrics in machine learning tasks. This observation was further validated through a correlation analysis between fidelity and utility results. The Pearson correlation coefficient between these properties for prostate cancer data was 0.185, for breast cancer data 0.033, and for cardiovascular diseases 0.34, all indicating weak relationships between these two properties.

This discrepancy can be attributed to the fact that fidelity metrics assess the alignment of individual attribute distributions between synthetic and real data. However, faithfully reproducing the marginal distributions of all attributes does not guarantee that the synthetic data will capture the joint distribution of the original dataset. This highlights the importance of selecting evaluation metrics tailored to the intended application of the synthetic dataset. For scenarios where the primary goal is to enhance machine learning model performance, as in cases of limited or low-quality training data, utility metrics provide a more relevant measure of synthetic data quality. Conversely, when the objective is to closely replicate the original attribute distributions for statistical analysis, attribute-specific tasks, or other purposes, fidelity metrics, such as CDF comparisons, are more appropriate for quality assessment.

For data quality comparison, we selected benchmarks such as CTGAN, CopulaGAN, and TVAE because these are well-established generative algorithms specifically tailored for tabular data, aligning closely with the focus of this study. The selection was carefully made to ensure relevance and provide meaningful insights into the performance of our methodology. However, tabular data generation is a rapidly evolving field, with innovative proposals for tabular data generation such as diffusion models or normalizing flows. Expanding the comparison to include these methods in future studies would be highly beneficial, offering deeper insights into the strengths and limitations of our UMAP-based approach and further contextualizing its performance within the broader landscape of generative models.

One key strength of the UMAP-based algorithm is its ability to generate large-scale synthetic datasets, even when the initial reference dataset is relatively small, as was the case with the prostate cancer and cardiovascular diseases datasets. This study successfully scaled fully synthetic datasets to over fifty times the size of the original dataset, illustrating the method's potential for substantial data augmentation.

Future work will focus on expanding the algorithm's applicability to larger, more diverse datasets. With bigger and more varied reference datasets, the algorithm can generate even larger synthetic datasets, as the synthetic data size scales exponentially with the reference dataset size. Additionally, we aim to explore the algorithm's performance with datasets containing a higher number of features, which could reveal how synthetic data quality varies across different attributes.

Although initially developed for healthcare applications, where data scarcity and privacy concerns are critical, the proposed algorithm is inherently generalizable and can be applied across various domains. The only requirement is the availability of a suitable reference dataset, making this methodology a versatile solution for addressing data limitations in diverse fields. Its reproducibility has already been demonstrated through promising results in three distinct healthcare domains: prostate cancer, breast cancer, and cardiovascular disease. Nevertheless, exploring its application beyond the healthcare domain remains an exciting avenue for future research, potentially broadening its impact and utility.

Regarding different domains, future research work could involve testing different dimensionality reduction tools. Although UMAP has shown excellent performance, specially for healthcare data, different fields could benefit from a different dimensionality reduction tool such as t-distributed stochastic neighbor embedding (t-SNE) or TriMap.

Although UMAP effectively transforms data with many attributes into a two-dimensional space, the iterative application of UMAP-based transformations for synthetic data generation can introduce computational challenges, particularly when scaling to datasets with a high number of features or large sample sizes. To address these challenges, future work should include a detailed computational efficiency analysis, assessing runtime, memory usage, and scalability. Such an analysis would help quantify the computational trade-offs involved and identify any potential bottlenecks.

Optimization strategies, such as parallelization or approximations to accelerate the UMAP computations, should also be considered to enhance scalability. For instance, distributed computing or GPU acceleration could significantly reduce runtime for high-dimensional data. Additionally, exploring dimensionality reduction techniques that complement or integrate with UMAP could provide alternative pathways for scaling the methodology to more complex datasets.

Author Contributions: Conceptualization, C.L. and C.A.; methodology, C.L.; software, C.L.; validation, C.L. and C.A.; formal analysis, C.L. and C.A.; investigation, C.L. and C.A.; resources, C.L.; data curation, C.L.; writing—original draft preparation, C.L.; writing—review and editing, C.L. and C.A.; visualization, C.L.; supervision, C.A.; project administration, C.A.; funding acquisition, C.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Union Horizon Europe R&I program through the project FLUTE (Grant agreement no.: 101095382). It is also co-funded by the European Union–Next Generation EU funds, through the I+D+i PID2021-122835OBC21 research project ARTIFACTS (MCIN/AEI/10.13039/501000011033) and FEDER "Una manera de hacer Europa".

Data Availability Statement: The original data presented in the study are openly available in the provided repository links.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

UMAP	Uniform Manifold Approximation and Projection
AI	Artificial Intelligence
PSA	Prostate-specific antigen level
t-SNE	t-distributed stochastic neighbor embedding
K-S	Kolmogorov–Smirnov
TSTR	Train on Synthetic, Test on Real
VAE	Variational Autoencoder
GAN	Generative Adversarial Network
CTGAN	Conditional Tabular Generative Adversarial Network
PI-CAI	Prostate Imaging and Cancer Analysis Information
CDF	Cumulative distribution function
SVM	Support vector machine
AUC	Area under the curve
ML	Machine learning

References

1. Feuerriegel, S.; Hartmann, J.; Janiesch, C.; Zschech, P. Generative ai. *Bus. Inf. Syst. Eng.* **2024**, *66*, 111–126. [CrossRef]
2. Ng, A.; Jordan, M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Adv. Neural Inf. Process. Syst.* **2001**, *14*, 841–848.
3. Shokouhifar, M.; Hasanvand, M.; Moharamkhani, E.; Werner, F. Ensemble Heuristic–Metaheuristic Feature Fusion Learning for Heart Disease Diagnosis Using Tabular Data. *Algorithms* **2024**, *17*, 34. [CrossRef]
4. Dixon, J.; Akinniyi, O.; Abdelhamid, A.; Saleh, G.A.; Rahman, M.M.; Khalifa, F. A hybrid learning-architecture for improved brain tumor recognition. *Algorithms* **2024**, *17*, 221. [CrossRef]
5. Topuz, K.; Zengul, F.D.; Dag, A.; Almehmi, A.; Yildirim, M.B. Predicting graft survival among kidney transplant recipients: A Bayesian decision support model. *Decis. Support Syst.* **2018**, *106*, 97–109. [CrossRef]
6. Cai, Q.; Wang, H.; Li, Z.; Liu, X. A survey on multimodal data-driven smart healthcare systems: Approaches and applications. *IEEE Access* **2019**, *7*, 133583–133599. [CrossRef]
7. Methaila, A.; Kansal, P.; Arya, H.; Kumar, P. Early heart disease prediction using data mining techniques. *Comput. Sci. Inf. Technol. J.* **2014**, *24*, 53–59.
8. Mosquera-Rojas, G.E.; Ouadah, C.; Hadadi, A.; Lalande, A.; Leclerc, S. Automatic Myocardium Segmentation in Delayed-Enhancement MRI with Pathology-Specific Data Augmentation and Deep Learning Architectures. *Algorithms* **2023**, *16*, 488. [CrossRef]
9. Shen, J.; Zhang, C.J.; Jiang, B.; Chen, J.; Song, J.; Liu, Z.; He, Z.; Wong, S.Y.; Fang, P.H.; Ming, W.K.; et al. Artificial intelligence versus clinicians in disease diagnosis: Systematic review. *JMIR Med. Inform.* **2019**, *7*, e10010. [CrossRef]
10. Xi, W.; Devineau, G.; Moutarde, F.; Yang, J. Generative model for skeletal human movements based on conditional DC-GAN applied to pseudo-images. *Algorithms* **2020**, *13*, 319. [CrossRef]
11. Preiksaitis, C.; Rose, C. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: Scoping review. *JMIR Med. Educ.* **2023**, *9*, e48785. [CrossRef]
12. Ayoub, M.; Ballout, A.A.; Zayek, R.A.; Ayoub, N.F. Mind+ Machine: ChatGPT as a Basic Clinical Decisions Support Tool. *Cureus* **2023**, *15*, e43690. [CrossRef]
13. Chen, Y.; Esmaeilzadeh, P. Generative AI in medical practice: In-depth exploration of privacy and security challenges. *J. Med. Internet Res.* **2024**, *26*, e53008. [CrossRef]
14. Toews, R. Forbes. Synthetic Data Is About to Transform Artificial Intelligence. 2023. Available online: https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/?sh=7d65c55d7523 (accessed on 20 June 2023).
15. Kingma, D.P. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
16. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
17. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning PMLR, Lille, France, 6–11 July 2015; pp. 2256–2265.
18. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
19. Tabak, E.G.; Turner, C.V. A family of nonparametric density estimation algorithms. *Commun. Pure Appl. Math.* **2013**, *66*, 145–164. [CrossRef]

20. Rezende, D.; Mohamed, S. Variational inference with normalizing flows. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 1530–1538.
21. Kim, D.K.; Ryu, D.; Lee, Y.; Choi, D.H. Generative models for tabular data: A review. *J. Mech. Sci. Technol.* **2024**, *38*, 4989–5005. [CrossRef]
22. Hernandez, M.; Epelde, G.; Alberdi, A.; Cilla, R.; Rankin, D. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* **2022**, *493*, 28–45. [CrossRef]
23. Coutinho-Almeida, J.a.; Rodrigues, P.P.; Cruz-Correia, R.J.a. GANs for Tabular Healthcare Data Generation: A Review on Utility and Privacy. In Proceedings of the Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, 11–13 October 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 282–291. [CrossRef]
24. Lázaro, C.; Angulo, C. Using UMAP for Partially Synthetic Healthcare Tabular Data Generation and Validation. *Sensors* **2024**, *24*, 7843. [CrossRef] [PubMed]
25. Hernandez, M.; Epelde, G.; Alberdi, A.; Cilla, R.; Rankin, D. Standardised metrics and methods for synthetic tabular data evaluation. *TechRxiv* **2021**. [CrossRef]
26. Yoon, J.; Mizrahi, M.; Ghalaty, N.F.; Jarvinen, T.; Ravi, A.S.; Brune, P.; Kong, F.; Anderson, D.; Lee, G.; Meir, A.; et al. EHR-Safe: Generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ Digit. Med.* **2023**, *6*, 141. [CrossRef]
27. Baowaly, M.K.; Lin, C.C.; Liu, C.L.; Chen, K.T. Synthesizing electronic health records using improved generative adversarial networks. *J. Am. Med. Inform. Assoc.* **2019**, *26*, 228–241. [CrossRef]
28. Bourou, S.; El Saer, A.; Velivassaki, T.H.; Voulkidis, A.; Zahariadis, T. A review of tabular data synthesis using GANs on an IDS dataset. *Information* **2021**, *12*, 375. [CrossRef]
29. Dankar, F.K.; Ibrahim, M.K.; Ismail, L. A multi-dimensional evaluation of synthetic data generators. *IEEE Access* **2022**, *10*, 11147–11158.
30. Esteban, C.; Hyland, S.L.; Rätsch, G. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv* **2017**, arXiv:1706.02633.
31. Mahendra, M.; Umesh, C.; Bej, S.; Schultz, K.; Wolkenhauer, O. Convex space learning for tabular synthetic data generation. *arXiv* **2024**, arXiv:2407.09789.
32. Rankin, D.; Black, M.; Bond, R.; Wallace, J.; Mulvenna, M.; Epelde, G. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR Med. Inform.* **2020**, *8*, e18910.
33. Wang, A.X.; Chukova, S.S.; Simpson, C.R.; Nguyen, B.P. Challenges and opportunities of generative models on tabular data. *Appl. Soft Comput.* **2024**, *166*, 112223. [CrossRef]
34. Mumuni, A.; Mumuni, F. Data augmentation: A comprehensive survey of modern approaches. *Array* **2022**, *16*, 100258. [CrossRef]
35. Micheletti, N.; Marchesi, R.; Kuo, N.I.H.; Barbieri, S.; Jurman, G.; Osmani, V. Generative AI Mitigates Representation Bias and Improves Model Fairness Through Synthetic Health Data. *medRxiv* **2023**. [CrossRef]
36. Caton, S.; Haas, C. Fairness in machine learning: A survey. *ACM Comput. Surv.* **2024**, *56*, 7. [CrossRef]
37. Wang, L.; Zhang, W.; He, X. Continuous patient-centric sequence generation via sequentially coupled adversarial learning. In Proceedings of the Database Systems for Advanced Applications: 24th International Conference, DASFAA 2019, Chiang Mai, Thailand, 22–25 April 2019; Proceedings, Part II 24; Springer: Berlin/Heidelberg, Germany, 2019; pp. 36–52.
38. Bietsch, D.; Stahlbock, R.; Voß, S. Synthetic Data as a Proxy for Real-World Electronic Health Records in the Patient Length of Stay Prediction. *Sustainability* **2023**, *15*, 13690. [CrossRef]
39. Raoof, I.; Gupta, M.K. A conditional input-based GAN for generating spatio-temporal motor imagery electroencephalograph data. *Neural Comput. Appl.* **2023**, *35*, 21841–21861. [CrossRef]
40. Titar, R.R.; Ramanathan, M. Variational autoencoders for generative modeling of drug dosing determinants in renal, hepatic, metabolic, and cardiac disease states. *Clin. Transl. Sci.* **2024**, *17*, e13872. [CrossRef] [PubMed]
41. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular data using Conditional GAN. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019.
42. Kamthe, S.; Assefa, S.; Deisenroth, M. Copula flows for synthetic data generation. *arXiv* **2021**, arXiv:2101.00598.
43. Nelsen, R.B. *An Introduction to Copulas*; Springer, New York, NY, USA, 2006 .
44. Liu, R.; Wei, J.; Liu, F.; Si, C.; Zhang, Y.; Rao, J.; Zheng, S.; Peng, D.; Yang, D.; Zhou, D.; et al. Best practices and lessons learned on synthetic data. In Proceedings of the First Conference on Language Modeling, Philadelphia, PA, USA, 7–9 October 2024.
45. Chauhan, P.; Bongo, L.A.; Pedersen, E. Ethical Challenges of Using Synthetic Data. In Proceedings of the AAAI Symposium Series, Arlington, VA, USA, 25–27 October 2023; Volume 1, pp. 133–134.
46. Vaiste, J. Ethical Implications of AI-Generated Synthetic Health Data. 2023. Available online: https://hal.science/hal-04216538v1/file/Ethical_implications_of_AI_generated_synthetic_medical_data-2.pdf (accessed on 15 December 2024).
47. Giuffrè, M.; Shung, D.L. Harnessing the power of synthetic data in healthcare: Innovation, application, and privacy. *NPJ Digit. Med.* **2023**, *6*, 186. [CrossRef] [PubMed]
48. Norori, N.; Hu, Q.; Aellen, F.M.; Faraci, F.D.; Tzovara, A. Addressing bias in big data and AI for health care: A call for open science. *Patterns* **2021**, *2*, 100347. [CrossRef] [PubMed]

49. Jordon, J.; Szpruch, L.; Houssiau, F.; Bottarelli, M.; Cherubin, G.; Maple, C.; Cohen, S.N.; Weller, A. Synthetic Data–what, why and how? *arXiv* **2022**, arXiv:2205.03257.
50. Beduschi, A. Synthetic data protection: Towards a paradigm change in data regulation? *Big Data Soc.* **2024**, *11*, 20539517241231277. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.