



## WP3 Synthetic image and data generation

---

### D 3.1 Structured synthetic health data generation

<https://www.fluteproject.eu/>



This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101095382. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the HaDEA. Neither the European Union nor the granting authority can be held responsible for them.

**Grant Agreement No.: 101095382**  
**Deliverable: D 3.1 Structured synthetic health data generation**

**Project Start Date:** 01/05/2023  
**Coordinator:** INRIA

**Duration:** 36 months

<b>Deliverable No:</b>	D 3.1
<b>WP No:</b>	3
<b>WP Leader:</b>	Angulo, Cecilio
<b>Due date:</b>	30/04/2025
<b>Delivery date:</b>	30/04/2025

**Dissemination Level:**

PU	Public Use	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

## DOCUMENT SUMMARY INFORMATION

<b>Project title:</b>	Federate Learning and mUlti-party computation Techniques for prostatE cancer
<b>Short project name:</b>	FLUTE
<b>Project No:</b>	101095382
<b>Call Identifier:</b>	HORIZON-HLTH-2022-IND-13
<b>Thematic Priority:</b>	HORIZON-HLTH-2022-IND-13
<b>Type of Action:</b>	HORIZON Research and Innovation Actions
<b>Start date of the project:</b>	01/05/2023
<b>Duration of the project:</b>	36 months
<b>Project website:</b>	<a href="https://www.fluteproject.eu/">https://www.fluteproject.eu/</a>

### D 3.1 Structured synthetic health data generation

<b>Work Package:</b>	WP3 Synthetic image and data generation
<b>Deliverable number:</b>	D 3.1
<b>Deliverable title:</b>	Structured synthetic health data generation
<b>Due date:</b>	30/04/2025
<b>Actual submission date:</b>	30/04/2025
<b>Editor:</b>	Angulo, Cecilio
<b>Authors:</b>	Cecilio Angulo, Carla Lázaro
<b>Dissemination Level:</b>	PU
<b>No. pages:</b>	76
<b>Authorized (date):</b>	25/04/2025
<b>Responsible person:</b>	INRIA
<b>Status:</b>	Plan/Draft/Working/ <i>Final</i> /Submitted/Approved

#### Revision history:

Version	Date	Author	Comment
v.0.1	15/11/2024		First outline, including a classification of the state-of-the-art
v.1.0	10/04/2025		First complete version
v.2.0	25/04/2025		Final version after internal review

#### Quality Control:

	Who	Date
Checked by internal reviewer	Carla Lázaro - UPC	11/04/2025
Checked by WP Leader	Angulo, Cecilio	17/04/2025
Checked by Project Technical Managers	GRAD, INRIA	25/04/2025
Checked by Project Coordinator	INRIA	25/04/2025

## COPYRIGHT

© Copyright by the FLUTE consortium, 2023-2026.

This document contains material, which is the copyright of FLUTE consortium members and the European Commission, and may not be reproduced or copied without permission, except as mandated by the European Commission Grant Agreement no. 101095382 for reviewing and dissemination purposes.

## ACKNOWLEDGEMENTS

FLUTE is a project that has received funding from the European Union's Horizon Europe research and innovation programme under Grant Agreement No 101095382. Please, for more information see <https://www.fluteproject.eu/>.

The partners in the project are Institut National de recherche en informatique et automatique, Fundación Centro Tecnológico de Telecomunicaciones de Galicia, Arteevo Technologies Ltd, Istituto Romagnolo per lo studio dei tumori dino amadori - IRST, Technovative Solutions Ltd, Time.Lex, Centre Hospitalier Universitaire de Liège, Universitat Politècnica de Catalunya, Fundacio Hospital Universitari Vall d'Hebron - Institut De Recerca, HL7 Europe Foundation, Quibim Sociedad Limitada . The content of this document is the result of the worked developed by the partners in the context of the project.

## DISCLAIMER

The content of the publication herein is the sole responsibility of the publishers and it does not necessarily represent the views expressed by the European Commission or its services. The information contained in this document is provided by the copyright holders "as is" and any express or implied warranties, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose are disclaimed. In no event shall the members of the FLUTE collaboration, including the copyright holders, or the European Commission be liable for any direct, indirect, incidental, special, exemplary, or consequential damages (including, but not limited to, procurement of substitute goods or services; loss of use, data, or profits; or business interruption) however caused and on any theory of liability, whether in contract, strict liability, or tort (including negligence or otherwise) arising in any way out of the use of the information contained in this document, even if advised of the possibility of such damage.

# Contents

<b>1</b>	<b>Introduction</b>	<b>10</b>
1.1	Specific project objective . . . . .	10
1.2	Relationship with other deliverables and work packages . . . . .	11
1.3	Ethical considerations for synthetic data in sensitive domains . . . . .	12
<b>2</b>	<b>State of the Art</b>	<b>14</b>
2.1	Generative AI market . . . . .	15
2.2	Definition of synthetic data . . . . .	16
2.3	Literature surveys . . . . .	17
2.4	Algorithms . . . . .	18
2.4.1	Variational Autoencoders (VAE) . . . . .	19
2.4.2	Generative Adversarial Networks (GANs) . . . . .	20
2.4.3	Diffusion Models (DM) . . . . .	21
2.4.4	Normalizing Flows (NF) . . . . .	21
2.5	Differential privacy . . . . .	22
2.6	Validation metrics . . . . .	23
2.6.1	Fidelity . . . . .	24
2.6.2	Utility . . . . .	26
2.6.3	Privacy . . . . .	28
2.6.4	Other properties – Diversity, Coverage, Density, Fairness . . . . .	29
2.7	Multimodal data . . . . .	30
2.8	Integration in federated learning . . . . .	31
<b>3</b>	<b>Datasets</b>	<b>32</b>
3.1	Prostate cancer datasets . . . . .	32
3.2	Breast cancer datasets . . . . .	33
3.3	Cardiovascular disease datasets . . . . .	34
<b>4</b>	<b>UMAP-PSDG Algorithm</b>	<b>35</b>
4.1	Motivation . . . . .	35
4.1.1	Fully and partially synthetic data . . . . .	36
4.1.2	Data visualization and dimensionality reduction tools . . . . .	36
4.2	Partially synthetic methodology proposal . . . . .	37
4.3	Results . . . . .	40
4.3.1	CIA synthetic data generation from PI-CAI . . . . .	40
4.3.2	BCNB synthetic data generation from BC-MLR . . . . .	40
4.3.3	Discussion of results . . . . .	42
<b>5</b>	<b>Iterative UMAP-FSDG</b>	<b>43</b>
5.1	Motivation . . . . .	43
5.2	Iterative UMAP FSDG methodology . . . . .	44
5.3	Results . . . . .	46
5.3.1	Applying the algorithm . . . . .	46
5.3.2	Discussion of results . . . . .	49

---

<b>6</b>	<b>Validation</b>	<b>49</b>
6.1	Algorithm UMAP-PSDG. Fidelity . . . . .	51
6.2	Algorithm UMAP-PSDG. Privacy and workflow strategies . . . . .	53
6.3	Algorithm Iterative UMAP-FSDG. Fidelity and utility . . . . .	56
6.3.1	Data fidelity . . . . .	56
6.3.2	Data utility . . . . .	56
6.3.3	Discussion of results . . . . .	60
<b>7</b>	<b>Conclusions and Future Work</b>	<b>62</b>

## List of Figures

1	Project flow and 2-stage clinical validation. . . . .	11
2	Synthetic image and data generation work package (framed in red) into the FLUTE work plan overview. . . . .	12
3	Synthetic data generation module (framed in red) into the FLUTE value chain and high level architecture. . . . .	13
4	Synthetic data generation module (framed in red) into the reference implementation architecture of FLUTE platform. . . . .	14
5	Projected replacement of real data with synthetic data in AI models over the next decade, from Gartner’s 2021 forecast [119]. . . . .	16
6	Representation of real (light gray) and artificial (dark gray) values for missing data (white) on fully synthetic, partially synthetic, and imputed data sets. . . . .	38
7	Synthetic data generation framework proposal. . . . .	39
8	Algorithm workflow . . . . .	45
9	Cardiovascular diseases data visualization (Ind-CardioDB). . . . .	46
10	Real samples from the incomplete database (light gray) and correctly imputed samples for different reliability values. . . . .	52
11	Number of synthetic samples per age. . . . .	53
12	Imputation performance of mean and $k$ -NN ( $k = 10$ ) vs. proposed methodology. . . . .	54
13	Workflow options according to data privacy. . . . .	55
14	Utility metrics (Accuracy, F1-score and AUC) for Random Forest model trained with real, synthetic and augmented data . . . . .	58
15	Utility metrics (Accuracy, F1-score and AUC) for Logistic Regression model trained with real, synthetic and augmented data . . . . .	59
16	Utility metrics (Accuracy, F1-score and AUC) for Support Vector Machine model trained with real, synthetic and augmented data . . . . .	61

---

## List of Tables

1	Quality metrics summary. . . . .	25
2	Feature correlation with target variable. . . . .	41
3	Mean feature disparity for prostate cancer reference data (PI-CAI) . . . . .	49
4	Percentage of correctly imputed samples with the proposed methodology, mean and $k$ -NN imputation. . . . .	55
5	K-S statistic for CDFs. Lower K-S statistic for each attribute in bold . . . . .	57

## Executive summary

This deliverable *D3.1. Structured synthetic health data generation* reports on the progress and the results of FLUTE's WP3 for Task T3.1. This includes the development of tools for structured synthetic data generation, and the results obtained from the statistical evaluation. A twin deliverable *D3.2* reports on image synthetic health data generation. A later deliverable *D3.3* will present the final results for multi-modal synthetic data generation being performed on the cloud-based FLUTE platform in a federated way. This federated synthetic data generation tool is a part of the services provided by the platform to researchers and innovators.

The main components of the deliverable are

- **UMAP-PSDG Algorithm.** We introduce a novel methodology for partially synthetic tabular data generation. Our approach is validated on prostate cancer and breast cancer datasets. An associated publication is available:

Lázaro, C., & Angulo, C. (2024). Using UMAP for Partially Synthetic Healthcare Tabular Data Generation and Validation. *Sensors*, 24(23), 7843.

<https://doi.org/10.3390/s24237843>

- **Iterative UMAP-FSDG Algorithm.** Building on the previously developed algorithm using data visualization techniques, this study extends to the generation of fully synthetic tabular healthcare data. This approach is successfully applied to three healthcare domains: prostate cancer, breast cancer, and cardiovascular disease. An associated publication is available:

Lázaro, C., & Angulo, C. (2024). Iterative Application of UMAP-Based Algorithms for Fully Synthetic Healthcare Tabular Data Generation. *Algorithms*, 17(12), 591.

<https://doi.org/10.3390/a17120591>

- **Validation.** 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. Moreover, the proposed methodology is designed not only to augment or complete data sets, but also to ensure data *privacy*.

As main results, it can be pointed out:

1. Iterative UMAP-based method outperforms GAN / VAE methods in terms of fidelity, particularly for breast cancer and cardiovascular disease data.
2. The synthetic and augmented datasets generated using the Iterative UMAP-based approach exhibit high utility (random forest, SVM, logistic regression).
3. The proposed Partially Synthetic Data Generation algorithm paves the way for a completely new line of research based on visualization tools.
4. We study the information workflow between separate data centres, taking into consideration privacy concerns, emphasizing the protection of sensitive information.

In conclusion, we make good progress towards achieving the KPI relevant to this part of the project and its associated specific project objective.

# 1 Introduction

The goal of the multi-disciplinary FLUTE project is to advance and scale up **data-driven healthcare** by developing novel methods for privacy-preserving cross-border utilization of data hubs.

The technical innovations of the project will be integrated in a privacy-enforcing platform that will provide innovators with a provenly secure environment for federated healthcare AI solution development, testing and deployment, including the integration of real world health data from the data hubs and the **generation and utilization of synthetic data (categorical, numerical and images)**.

The project will integrate the FLUTE platform with **health data hubs located in three different countries**, use their data to develop a novel federated AI toolset for **diagnosis of clinically significant prostate cancer**, avoiding unnecessary biopsies, thus improving the welfare of patients and significantly reducing the associated costs.

## 1.1 Specific project objective

FLUTE's Project Objective 3 (PO3) refers to **research and develop novel methods for the generation of multi-modal synthetic health data**. Using multimodal EHR data from hospitals, biopsy images and multiparametric MRI and measured variables from health data lakes, develop novel methods of synthetic multi-modal health data generation. Going beyond the current methods based on Generative Adversarial Networks (GANs) that generate synthetic data by converting real-world data into 2-D objects, the project application built on developing new methods that employ new structures and architectures. Moreover, it was proposed to develop new models based on GANs and Diffusion Models to produce bi- and multi-parametric full 3D images that are consistent across contrasts. Usage of the novel synthetic data will be for augmentation of Federated Learning training data and model conditioning.

Deliverable *D3.1. Structured synthetic health data generation* contains the results and summarizes the developments of Task T3.1. This includes the development of tools for structured synthetic data generation, and the results obtained from the statistical evaluation. Hence, research and innovation project objectives will be focused on:

PO3.1 -D3.1 **Algorithms for structured synthetic health data generation**, mainly tabular data.

PO3.2 - D3.1 **Metrics for evaluation** of the generated structured synthetic health data, mainly statistical evaluation, but also privacy considerations.

PO3.3 - D3.2 **Definition of multi-modal structures** to make compatible generated structured synthetic health data with any other kind of data, specially images.

PO3.4 - D3.3 **Integration into federated learning mechanisms** of the generated structured synthetic health data to condition learning processes, mainly in the form of data augmentation.

**Outcomes and measurable results** Synthetic data generation strongly depends on accessing real world data from the hospitals through the FLUTE platform. Hence, results from this WP rely on the defined 2-stage project flow (see Figure 1): before FLUTE, and after FLUTE availability. In the first stage accessed data is coming from public repositories. In a second step, developed algorithms can be validated on data provided by the clinical partners.

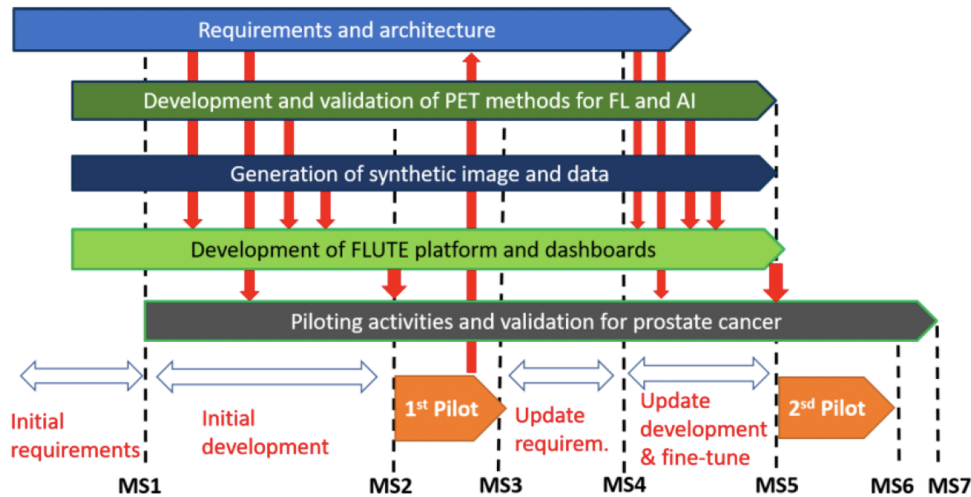


Figure 1: Project flow and 2-stage clinical validation.

**KPI and validation** As it was previously mentioned, algorithm development and results validation are implemented in public repositories. Extrapolation to data from the clinical partners should be straightforward. Moreover, robustness of the methods is tested on data coming from related but near clinical domains. Outcomes and results can be validated.

1. Synthetic data generator models are developed based on multi-modal health data;
2. Synthetic databases are generated using the generator models subjected to statistical tests;
3. Two sets of synthetic data will be integrated in FL in Task T3.3 by combining local models and synthetic data, and validated using sensitivity, specificity, precision, F1-score and AUC metrics, with better results than in similar approaches that do not use synthetic data.

## 1.2 Relationship with other deliverables and work packages

The deliverable *D3.1 Structured synthetic health data generation* will contain the results and summarize the developments of Task T3.1. This includes the development of tools for structured synthetic data generation, and the results obtained from the statistical evaluation.

The task *T3.1: Research and development of structured synthetic health data generation* is mainly related with project objective PO3, but also with project objectives PO2 and PO4, that is,

- PO2. Achieve high **scalability** of Federated Learning while maintaining state of the art levels of privacy and AI model performance.

- PO4. Design and develop the FLUTE platform.

This task will develop algorithms for the generation of synthetic clinical data and quantitative imaging biomarkers. The proposed algorithms will take into account the nature of the data as well as the FL platform defined and developed in T4.2. Going beyond the current Generative Adversarial Network (GAN) based methods that generate synthetic data by converting real world data into 2-D objects, new methods are developed that employ autoencoder structures, Visualization Algorithms, or Diffusion Models architectures. Statistical tools are employed for quantifying the synthetic data performance. Synthetic image and data generation work package is shown in Figure 2 framed in red into the FLUTE work plan overview.

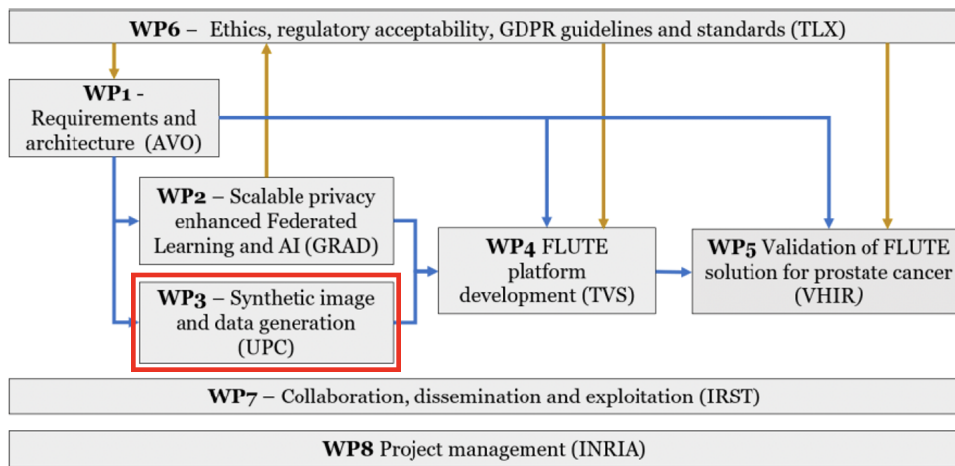


Figure 2: Synthetic image and data generation work package (framed in red) into the FLUTE work plan overview.

Synthetic data generation will be performed on the cloud-based FLUTE platform, as part of the services provided by the platform to researchers and innovators. Figure 3 shows the FLUTE value chain and high-level architecture. Innovators will be able to use the synthetic data for model architecture evaluation and for initial coarse training of models under development, drastically reducing the need for real-world data in the development and experimentation with new AI models, thus fostering rapid digital healthcare innovation. The reference implementation architecture is shown in Figure 4.

### 1.3 Ethical considerations for synthetic data in sensitive domains

Synthetic data offers significant opportunities to address biases in healthcare datasets and alleviate privacy concerns. However, its use raises important ethical challenges that require careful consideration, including issues related to fairness, privacy, and potential misuse [92]. These challenges highlight the need for a nuanced understanding of the limitations and implications of synthetic data in sensitive domains such as healthcare [26].

Fairness remains a central concern when using synthetic data. Despite its potential, synthetic data is not inherently unbiased. The generation process relies on real datasets, which often contain patterns and biases embedded in the original data [57]. Consequently, any bias present in the source data is likely to be replicated, or even amplified, in the synthetic dataset.

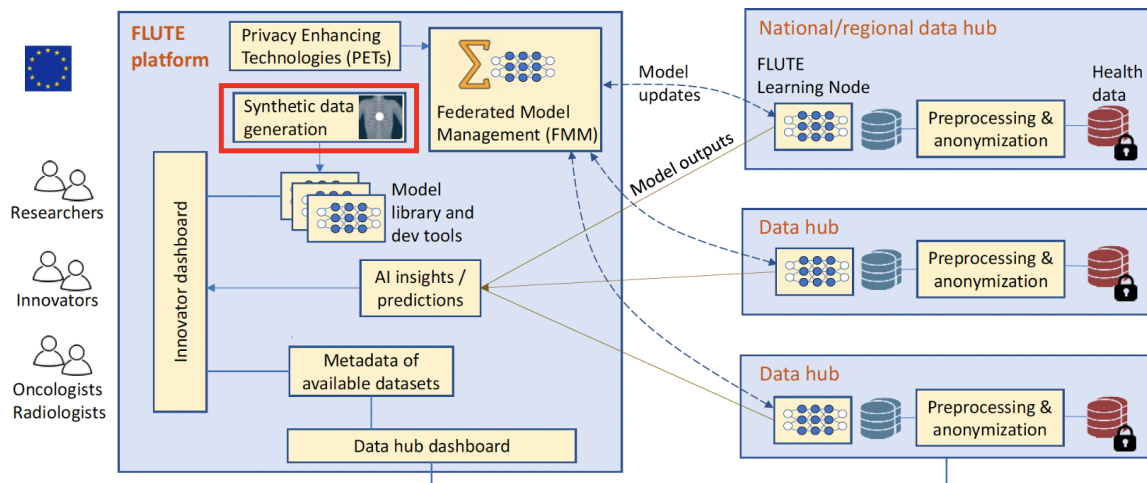


Figure 3: Synthetic data generation module (framed in red) into the FLUTE value chain and high level architecture.

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.

Privacy considerations also require significant attention. Although synthetic data is 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 is not automatically private” [79]. 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 also elements to be considered 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 its limitations and make informed decisions about its use. These features are considered in WP6.

Best practices for responsible synthetic data usage include regular bias assessments to identify and mitigate any disparities [57]. 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 [13]. 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 worth noting that synthetic data cannot fully replace real-world datasets [79]. By design, synthetic data introduces some level of distortion to ensure privacy or mitigate biases. These distortions, while necessary, may affect the validity of downstream analyses and

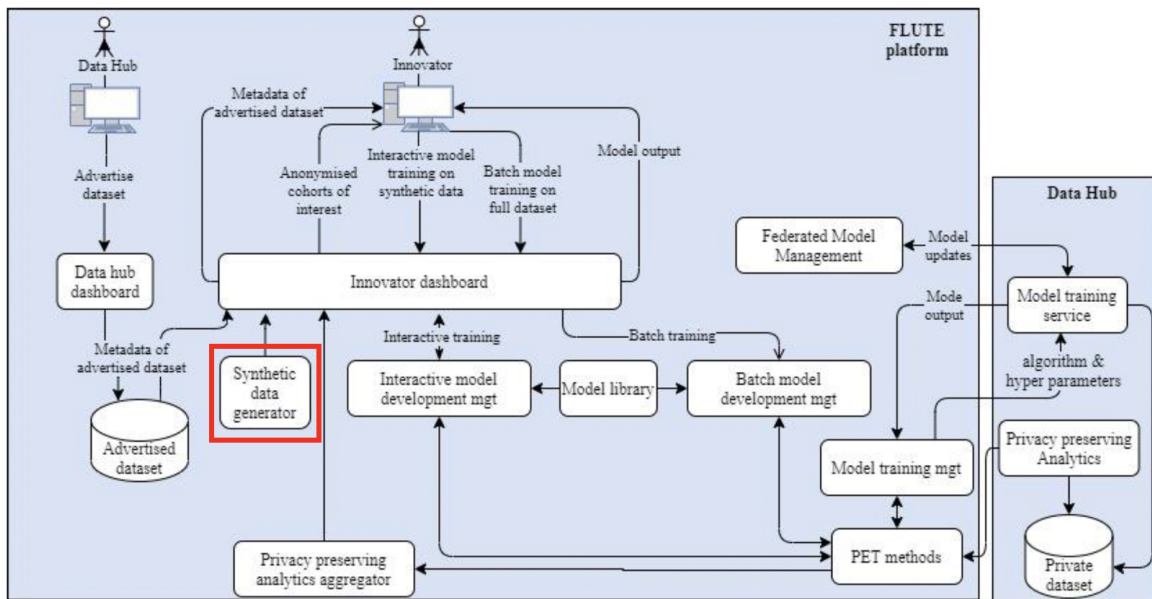


Figure 4: Synthetic data generation module (framed in red) into the reference implementation architecture of FLUTE platform.

model training. As such, synthetic data should primarily serve as a complementary resource to accelerate the research and innovation 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.

## 2 State of the Art

This section is devoted to the state of the art about novel methods of synthetic healthcare data generation. In particular, focus will be posed on (i) tabular synthetic data generation and (ii) validation tools of the synthetic generated data.

As it was written in the project application, health data hubs may contain data in different formats. The multicategory logical, date, ordinal and continuous data from the hospitals are typically in the form of EHRs, while oncology data lakes provide biopsy and biparametric (bp) and multiparametric (mp) MRI images and measured variables of the tumors, such as size, position and volume. The EHRs and parameters extracted from MRI are usually converted into 2-D objects that can be processed by Generative Adversarial Networks (GANs) to generate synthetic data: specifically for cancer patients [60], or medGAN [31] and MC-medGAN [23] for EHRs-based secure federated AI model development. In FLUTE, these methods are extended to generate novel information. Moreover, novel approaches are explored by constructing datasets in the form of 1-D objects [28]. For bpMRI and mpMRI, the existing models are limited to the generation of 2D slices or small patches due to memory constraints [35]. In FLUTE, new models based on GANs, Autoencoders and Diffusion Models are developed to generate full 3D volumes. Two approaches are considered: (a) based on the synthesis of a single parametric MRI and image-to-image translation, generate the other MRI modalities; (b) generate all modalities in a single stage while ensuring the consistency

of the parametric information. Novel uses of synthetic data in FL will be applied based on recent work [45] with insertion in FL considered to be a form of data augmentation for training local models on local and synthetic global data.

## 2.1 Generative AI market

Generative AI refers to advanced computational techniques that can produce novel, meaningful content—such as text, images, or audio—based on patterns from training data [50]. This rapidly evolving field has produced well-known applications such as DALL-E 2, GPT-4, and Copilot that are reshaping communication, creativity, and professional workflows across many industries.

At its core, generative AI relies on generative modeling, a class of machine learning distinct from discriminative models. While discriminative models are commonly employed in tasks like classification and prediction [112], generative models seek to understand the underlying data distribution and can create new data points similar to the training data. This ability to simulate new data with realistic attributes has paved the way for a range of practical applications.

Generative AI has already begun to revolutionize various sectors by enabling solutions that address real-world challenges. Applications range from content generation, such as writing search-engine-optimized text, to code generation through tools like Copilot, which simplifies software development. These innovations are driving new levels of efficiency and creativity, fostering advancements in industries from marketing [78] to healthcare [2].

In the medical field, generative AI holds significant potential for improving diagnostics [132], research [117] and treatment planning [9]. By generating synthetic patient data, augmenting rare disease studies, or enhancing drug discovery pipelines, generative AI could help detect patterns, predict disease progression, and support clinical decision-making [29]. However, the sensitive nature of medical data introduces challenges related to the protection of health information, underscoring the need for privacy-conscious AI development.

To address these concerns, strategies for managing generative AI risks include developing robust risk assessment protocols, establishing trustworthiness and responsibility metrics, and implementing continuous model monitoring [82, 107, 152]. However, these safeguards alone may not be sufficient. Comprehensive governance frameworks and updates to privacy laws are essential to ensure that generative AI evolves responsibly and securely.

The ethical and regulatory landscape also poses significant hurdles for companies operating in the generative AI sector. With data privacy regulations such as the European Union's General Data Protection Regulation (GDPR) [75] and the California Consumer Privacy Act (CCPA) [108] imposing strict requirements on the collection and processing of personal data, businesses must carefully navigate legal obligations. This is a topic being analysed in WP6.

Given the strict limitations on accessing and using real data in many industries, especially those involving sensitive information like healthcare and finance, synthetic data provides a way to bypass these obstacles. Real data is often restricted due to privacy concerns, limited availability, or imbalances in class representation, which can hinder AI model performance. Synthetic data addresses these issues by enabling the creation of large, diverse datasets that

maintain the statistical properties of real data without exposing personal information, and eventually, improving the model performance.

Both industry and academia have recognized the transformative power of the generative AI movement. Companies in the generative AI space are capitalizing on this potential. For instance, Mistral AI, a startup focused on developing cutting-edge open-source generative AI models, was valued at 260 million within just four weeks of its founding, raising 113 million in seed funding to compete with industry leaders like OpenAI, creators of ChatGPT [20]. Other companies, such as Gretel.ai and Mostly AI, are leading the way in synthetic data generation, helping businesses across industries to leverage privacy-compliant synthetic datasets for training AI models. These companies showcase the growing commercial demand for synthetic data solutions and highlight how this technology is driving innovation.

Looking forward, organizations like Gartner predict that synthetic data will completely eclipse real data in AI models by 2030 (see Figure 5), as privacy and security concerns continue to grow [141]. Since synthetic data enables AI models to perform at higher levels without risking data privacy violations, it is poised to become a fundamental pillar in the expansion of the generative AI market, helping organizations innovate while staying compliant with regulatory frameworks.

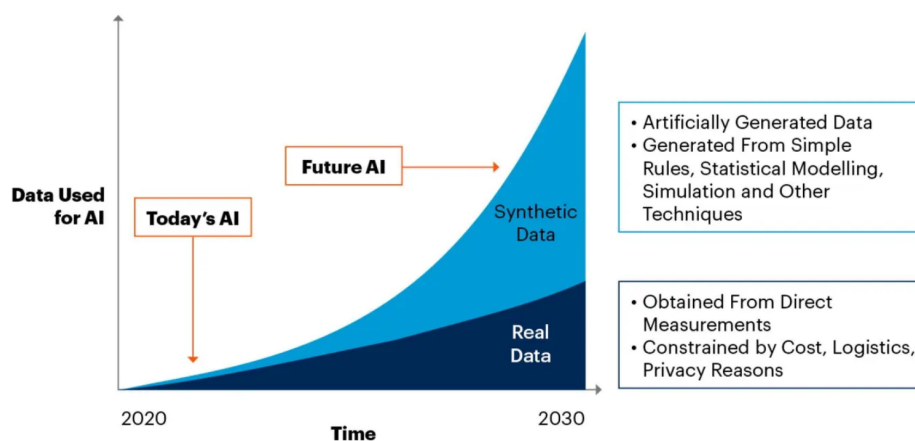


Figure 5: Projected replacement of real data with synthetic data in AI models over the next decade, from Gartner’s 2021 forecast [119].

## 2.2 Definition of synthetic data

The term *synthetic data* (SD) refers to data that is artificially generated to replicate the statistical properties of real data. Its primary objective is to maintain the quality of the real data while mitigating privacy risks by preventing the disclosure of sensitive information about individuals [11].

Several perspectives on the definition of SD have been proposed in the literature, reflecting its versatility and adaptability to various contexts and objectives. Depending on the purpose for which SD is defined and used, the following classification can be identified:

- **Replicating Original Data:** El Emam et al. [45] defines synthetic data as “data generated from real data that retains the same statistical properties as the original.” This per-

spective emphasizes the capacity of SD to act as a substitute for real datasets, enabling comparable analytical outcomes while safeguarding privacy.

- **Replicating Complex Patterns:** Synthetic data is described in [6, 52] as “data generated by generative models that learn the underlying patterns and distributions of real-world data.” This definition underscores the pivotal role of algorithms, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), in creating synthetic datasets.
- **Task-Oriented Data:** Jordon et al. [79] defines synthetic data as “data generated using a purpose built mathematical model or algorithm to solve a set of data science tasks.” This perspective highlights the task-specific nature of SD, tailored to address challenges such as predictive modeling or data augmentation.
- **Privacy-Preserving Characteristics:** Giomi et al. [56] characterize synthetic data as a “means of sharing sensitive information while preserving the statistical properties of the original data without disclosing individual-level information”. This definition emphasizes its application in privacy-sensitive domains, particularly where regulatory compliance is critical.

Since we are considering the use of synthetic data in all their dimensions, this deliverable will consider all the definitions equally valid.

## 2.3 Literature surveys

The scarcity of patient data for research purposes, largely due to privacy concerns and regulatory restrictions, has led to the increased use of synthetic data as a privacy-preserving alternative. Synthetic data has gained significant attention in research and academic domains. For instance, Murtaza et al. [110] conducted a comprehensive narrative review of 70 peer-reviewed studies, evaluating different techniques for generating privacy-preserving synthetic medical data. Their review highlights the utility of synthetic datasets in various research applications, innovation, academics, and testing. However, the focus on longitudinal synthetic data seems deficient. Moreover, a unified metric for generic quality assessment of synthetic data is lacking.

In [51], the state-of-the-art concerning Generative Adversarial Networks (GANs) for synthetic data generation was analyzed. The authors reviewed literature from major bibliographic databases—Web of Science, Scopus, IEEE Xplore, and ACM Digital Library—by focusing on studies published after 2010 that explore GANs in the context of synthetic data generation. Notably, one of the top-cited studies within this body of literature applied GANs to generate synthetic medical images for liver lesion classification [53]. Despite the heavy focus on image generation, the review emphasizes the need for further research into synthetic tabular data generation, while also noting that the evaluation of synthetic data quality remains subjective, heavily depending on the specific task and domain.

The research gap in synthetic tabular data generation is further discussed in [33], where GAN-based methods are specifically reviewed for generating tabular healthcare data. Published in 2022, their review covered 22 studies between 2017 and 2022 and categorized these papers into three key themes: utility, privacy, and clinical relevance. They examined meth-

ods used to measure the quality of generated data, privacy safeguards, and evaluations conducted by healthcare professionals. Their findings suggest that while GANs show promise in this area, the volume of research remains limited, and there is a pressing need for further development.

A more detailed and systematic review of synthetic data generation for tabular health records is offered by Hernández et al. [71]. The study focused on approaches used in healthcare for synthetic tabular data generation, particularly examining the contributions of GAN-based methods.

Several use cases of synthetic data in healthcare were explored in [59]. Through a review of literature across PubMed, Scopus, and Google Scholar, they identified seven core applications: simulation and prediction research, hypothesis and algorithm testing, epidemiology/public health, health IT development, education and training, public dataset releases, and data linkage. Their review demonstrates the breadth of synthetic data's potential impact on healthcare research and practice.

Lastly, a review [83] is published in 2024 focused on generative models for tabular data. The authors systematically categorized studies based on objectives such as differential privacy, data distribution modeling, and data augmentation. They presented evaluation methods specific to these objectives, including metrics like correlation coefficient,  $\epsilon$ -loss, F1-score, rooted mean squared error, Kolmogorov-Smirnov test, and Kullback–Leibler divergence, providing a comprehensive overview of performance comparisons across the reviewed studies.

## 2.4 Algorithms

The generation of synthetic data serves multiple purposes, such as regularizing machine learning classifiers through data augmentation, anonymizing datasets, and supporting semi-supervised and self-supervised learning frameworks. In imbalanced learning contexts, where there is a significant disparity in class frequencies, synthetic data is essential to improve model performance by balancing class representation and mitigating bias.

One of the earliest algorithms for synthetic data generation is SMOTE (Synthetic Minority Over-sampling Technique) [27], introduced in 2002 by Chawla et al. SMOTE addresses the problem of imbalanced datasets by generating synthetic samples for the minority class. This algorithm generates new data points by interpolating between neighboring minority class instances, offering a more effective approach than simple random oversampling (ROS). Over the years, several SMOTE variants have emerged to refine the technique. Borderline-SMOTE [66] and Safe-Level-SMOTE [21] focus on generating samples in safe regions, while ADASYN [68] (Adaptive Synthetic Sampling) adjusts the number of synthetic samples based on the density of minority class instances, improving the learning process.

In addition to SMOTE-based approaches, other classical data generation techniques, such as rotation, scaling, and perturbation, have been widely employed for data augmentation, especially in scenarios with limited or low-quality data. These techniques aim to create new data points by modifying the existing dataset, helping models generalize better when data is scarce or of poor quality.

Early work in synthetic data generation also relied on stochastic models, such as Gaussian

Processes (GP) and Bayesian Networks (BN). Gaussian Processes extend multivariate Gaussian distributions to an infinite number of variables, offering a probabilistic approach to modeling complex data [98]. Bayesian Networks, on the other hand, use a graphical structure to represent conditional independence between features, combined with local conditional probability distributions, to encode the joint distribution of a dataset [54]. These stochastic methods laid the groundwork for generative approaches by offering mathematically grounded ways to model data distributions.

In the 2000s, research on generative models largely focused on machine learning techniques such as decision trees (DT), random forests (RF), and support vector machines (k-s) [83]. These machine learning models offered significant advantages over conventional stochastic methods, particularly in handling datasets with numerous attributes. Their ability to perform well in complex scenarios contributed to their widespread adoption for generating tabular data.

While many of these traditional methods are still in use today, the rise of deep learning in the 2010s marked a major shift in synthetic data generation. Deep learning models brought a new level of sophistication to this field, outperforming machine learning-based approaches in terms of generative performance. By leveraging deep learning's capacity to analyze complex, non-linear relationships between attributes, these models enabled the generation of highly realistic tabular data that closely reflects the distribution of the original dataset.

Among the most prominent deep learning techniques are Variational Autoencoders (VAE), Generative Adversarial Networks (GAN), Diffusion Models (DM) and Normalization Flows (NF), which represent the cutting-edge approaches to synthetic data generation today.

### 2.4.1 Variational Autoencoders (VAE)

Autoencoder neural networks are composed of two primary components: an encoder and a decoder. The encoder transforms input data into a latent space representation, while the decoder attempts to reconstruct the input from this representation with minimal error. In this way, autoencoders learn efficient representations of the data. However, traditional autoencoders map the input to a deterministic point in the latent space, limiting their ability to generate new samples.

Variational Autoencoders (VAEs), introduced by Kingma et al. in 2013 [84], extend the autoencoder framework by incorporating a probabilistic element to the latent space. Instead of mapping each input to a single point, VAEs map the input to a distribution—typically Gaussian—defined by two parameters: a mean and a variance. This allows for sampling from the latent space, enabling the generation of new data points that follow the underlying distribution of the input data.

While VAEs have shown remarkable success in generating images, adapting them for tabular data generation presents unique challenges. Tabular data often contains both discrete and continuous features, with some columns being highly imbalanced or containing multimodal distributions. This poses difficulties for traditional VAEs, which assume a Gaussian latent space, as it may not be well-suited for representing the complexities of tabular data.

To address these challenges, several modifications of the original VAE architecture have been proposed for tabular data generation. One notable approach is Tab-VAE [139], specifically

designed to handle the intricacies of tabular data. Tab-VAE modifies the VAE framework to better model both discrete and continuous columns while addressing issues such as the imbalance of categorical variables and the non-Gaussian nature of continuous columns.

Further advancements in VAE-based models for tabular data include VAE-GMM [7], which integrates a Bayesian Gaussian Mixture Model (BGM) within the VAE architecture. This model relaxes the assumption of a strictly Gaussian latent space, allowing for a more flexible and accurate representation of the underlying data distribution.

Another recent development is OVAE (Oblivious VAE) [144], which introduces differentiable oblivious decision trees (DODTs) into the VAE framework. This hybrid approach leverages the strengths of both VAEs and decision trees to improve the quality of generated synthetic tabular data.

Although VAEs have been explored for synthetic tabular data generation, much of the recent research has focused on Generative Adversarial Networks (GANs).

#### 2.4.2 Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs) were first introduced by Ian Goodfellow et al. in 2014 [61], being a more advanced generative model than the VAE model. GANs consist of two neural networks: the generator and the discriminator, both trained in a competitive setting. The generator is responsible for creating synthetic data that resembles real-world data, while the discriminator role is to assess the authenticity of this data. The two networks are trained simultaneously, with the generator aiming to deceive the discriminator into classifying its output as real data, and the discriminator learning to distinguish between real and generated data. This adversarial training process allows GANs to generate highly realistic data, making them particularly effective for complex data generation tasks.

Although GANs initially found success in computer vision, particularly for generating and synthesizing image data, recent efforts have adapted them for tabular data generation. One notable example is CTGAN [154], which modifies the GAN architecture to handle the specific challenges of tabular data, such as the coexistence of continuous and categorical variables. Another innovative example is PATE-GAN [80], which combines the Private Aggregation of Teacher Ensembles (PATE) framework with GANs to enforce privacy preservation during the data generation process. PATE-GAN approach ensures tight privacy guarantees by limiting the influence of individual samples on the model, thereby enhancing privacy without compromising performance.

In the healthcare domain, GANs have been applied to generate synthetic patient data. For instance, MedGAN [32] is designed to generate high-dimensional discrete variables, such as binary and count features, by combining an autoencoder with a GAN. This method is especially useful for simulating patient records with mixed data types. CorrGAN [116] builds on the MedGAN architecture by adding a correlation preservation mechanism, ensuring that the generated data maintains the original data attribute correlations. Further improvements to the GAN-based approach for medical data generation were proposed by Camino et al. [24], who introduced the Wasserstein GAN (WGAN) [8] with gradient penalty to stabilize training and improve the quality of generated data.

### 2.4.3 Diffusion Models (DM)

Diffusion models [135, 74] are generative models that employ a stochastic process to iteratively transform noise into meaningful data samples. These models have gained significant attention due to their remarkable performance across various generative tasks, particularly in text-to-image synthesis. Unlike GANs, which rely on adversarial training, diffusion models use a progressive denoising process, making them more stable and yielding higher-quality results in many cases, especially for complex data generation.

The impressive success of diffusion models in image synthesis has sparked interest in applying them to the generation of tabular data. One of the most promising examples is TabD-DPM [86], a recent diffusion-based model designed specifically for tabular data. TabDDPM uses multilayer perceptrons as its backbone and excels at learning unbalanced distributions from the training data. By employing Markov transition functions, referred to as diffusion kernels, for both continuous and discrete features, TabDDPM marks significant progress in tabular data synthesis. Its ability to capture complex data relationships and handle imbalanced distributions underscores the potential of diffusion models as a powerful solution in this domain.

Another application of diffusion models can be found in the healthcare sector, where MedDiff [69] was introduced as a generative model for electronic health records (EHRs). MedDiff represents the first successful implementation of diffusion models for medical data, demonstrating their versatility beyond image-based tasks. In the financial domain, diffusion models have also shown promise. FinDiff [128], for instance, is designed to generate synthetic financial data. It employs embedding encodings to handle the mixed modalities of financial datasets, which often include both categorical and numeric attributes, providing a tailored solution for this specific context.

A final noteworthy development is the introduction of a fair diffusion model by Yang et al. [159]. This model is designed to address biases present in sensitive attributes of training data, making it suitable for generating balanced datasets that mitigate issues of fairness. Through empirical evaluation, the model demonstrates its ability to reduce class imbalance in the generated data while maintaining the quality of the synthetic samples.

### 2.4.4 Normalizing Flows (NF)

Normalizing Flows [138, 125] are a family of generative models that enable efficient sampling and density evaluation by transforming a simple probability distribution, like a standard normal, into a more complex one through a series of invertible and differentiable mappings [85]. The key advantage of NFs is their ability to exactly compute both the likelihood of the data and generate new samples by applying the inverse transformations, while accounting for changes in volume through the Jacobian of each transformation.

While Normalizing Flows have been successfully applied in areas such as image, video, audio, and graph generation, their use for tabular data is still developing. Recent work has proposed frameworks such as a differentially private normalizing flow for heterogeneous tabular data [90], and CeFlow [40], which utilizes normalizing flows to model both continuous and categorical features effectively.

## 2.5 Differential privacy

According to the FLUTE's project objective *PO3.4*, to be mainly developed in the Deliverable *D3.3*, the generated structured synthetic health data conditioning learning processes will be integrated into a federated learning mechanism. Unlike basic anonymization, differential privacy is a very popular mathematical framework in federated setups that ensures the privacy of individuals in a dataset by adding carefully calibrated noise. Hence, it is ideal for high-stakes uses like healthcare, where both accuracy and confidentiality matter.

As it is pointed out in [41], differential privacy is defined as a privacy guarantee that ensures the addition or removal of a single item in a database does not significantly impact the outcome of any analysis performed on that data. This guarantee is formalized through a randomized function  $\mathcal{K}$ , called *mechanism*, which provides  $\epsilon$ -differential privacy. Specifically, for any two *adjacent* datasets,  $D_1$  and  $D_2$ , that differ by at most one element, and for any subset  $S$  within the range of  $\mathcal{K}$ , the following condition must hold:

$$Pr[\mathcal{K}(D_1) \in S] \leq \exp(\epsilon) \times Pr[\mathcal{K}(D_2) \in S] \quad (1)$$

Multiple protocols for introducing differential privacy have been proposed over the last years [42]. One commonly used mechanism to achieve differential privacy, as introduced in [43], involves adding random noise to the data. This approach allows the output of a function to be perturbed in such a way that it remains accurate at an aggregate level while protecting individual privacy. Consider a function  $f$  applied to a database  $X$ , which produces a true result  $f(X)$ . To generate a private response, a noise mechanism  $\mathcal{Q}$  is applied, adding carefully calibrated random noise to the true answer:

$$f(X) + (\text{Lap}(\Delta f/\epsilon))^k \quad (2)$$

Here, noise is drawn independently of a Laplace distribution,  $\text{Lap}(\Delta f/\epsilon)$ , for each of the  $k$  components of  $f(X)$ .

The parameter  $\Delta f$ , known as the *sensitivity* of the function  $f$ , measures the maximum change in  $f$ 's output when a single database element is altered. Mathematically, for a function  $f : \mathcal{D} \rightarrow \mathcal{R}^k$ , the sensitivity is defined as:

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \quad (3)$$

where  $D_1$  and  $D_2$  are datasets that differ by at most one element. It is important to note that sensitivity is an inherent property of the function  $f$  and does not depend on the actual data in the database.

The Laplace distribution  $\text{Lap}(\Delta f/\epsilon)$  used for adding noise is a scaled symmetric exponential distribution with a standard deviation of  $\sqrt{2}\Delta f/\epsilon$ . By adjusting the amount of noise according to  $\epsilon$ , one can control the balance between privacy and accuracy.

Furthermore,  $\epsilon$ -differential privacy can be maintained for any sequence of functions  $f_1, \dots, f_d$  by applying the noise mechanism  $\mathcal{Q}$  with a noise distribution scaled to  $\text{Lap}(\sum_i \Delta f_i/\epsilon)$ . This approach allows for the cumulative privacy guarantees over multiple queries, ensuring that the overall privacy loss remains within the desired bounds.

Differential privacy has been widely adopted across various fields to safeguard individual information, particularly in domains where data sensitivity is paramount. One such area is healthcare, where the confidentiality of health records is of utmost importance.

In the realm of biomedical data, differential privacy has been integrated into generative models to protect patient privacy while still allowing useful data analysis. For instance, [12] explore the use of differential privacy within a Generative Adversarial Network (GAN) framework. Their study evaluates how differentially private GANs can generate synthetic biomedical data that remains suitable for valid reanalysis while minimizing the risk of exposing participant information.

Similarly, Lee et al. [89] propose a different approach by using a generative autoencoder to synthesize electronic health records. They apply differential privacy techniques to this model to prevent privacy leakage of sensitive patient data, thereby enabling the secure sharing of synthetic health records for research purposes.

The widespread adoption of differential privacy has led to a distinction between two main models: traditional differential privacy, also known as centralized differential privacy (CDP), and local differential privacy (LDP) [156]. In the centralized differential privacy model, a service provider collects raw user data and then adds noise to the aggregated data before releasing it as statistical information to the public. This approach, however, poses significant challenges in ensuring the privacy of individual users' data.

To address these privacy concerns, the concept of local differential privacy (LDP) has been introduced. Under LDP, data is perturbed locally on users' devices before it is transmitted to a central server. This means that only the randomized, privacy-preserving data is shared, reducing the risk of privacy breaches at the server level.

Despite its advantages in enhancing privacy, local differential privacy has certain drawbacks. In particular, the noise added to protect privacy can be substantial when applied to the entire dataset, which often leads to a significant reduction in the utility of the query responses compared to the centralized model. Nonetheless, LDP has been successfully implemented in various applications over recent years, such as those described in [46, 36], demonstrating its practical relevance despite the trade-offs involved.

Furthermore, the application of differential privacy extends to Local Differential Privacy (LDP) within the healthcare domain. [64] utilize LDP in conjunction with a Generative Adversarial Network to protect training data composed of medical records. This method ensures that the privacy of individuals is maintained even during the model training process, where data vulnerability can be particularly high.

## 2.6 Validation 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 receives less attention compared to synthetic image data. For example, [72] 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.

Historically, the quality of synthetic data was predominantly measured by its **fidelity or resemblance**, focusing on statistical similarity between the synthetic data and the original dataset [110]. Over time, however, the emphasis has broadened to include metrics that assess how well synthetic data performs in its intended use, marking a shift from mere resemblance to task-specific evaluation. As **privacy** concerns have grown, new metrics have been introduced to measure the risk of identity disclosure in synthetic datasets [113]. Privacy evaluation has also expanded to include differential privacy mechanisms, which provide quantitative ways to measure the risk of identity disclosure while ensuring data protection [147]. Furthermore, recent developments have incorporated more specific metrics, such as diversity, coverage, density, and fairness, to assess both the **utility** and ethical considerations of the synthetic data.

When selecting validation metrics, the intended application of the synthetic data must be carefully considered. For instance, if the goal is to augment sample sizes in clinical trials, validity could be assessed by comparing the synthetic data to the full dataset. In cases where synthetic data serves as a privacy-preserving substitute for real data, its validity might be evaluated through statistical distribution comparisons or machine learning performance metrics, comparing results obtained from synthetic data with those derived from real data.

Hence, key properties used in the evaluation of synthetic data include:

- **Fidelity / Resemblance:** Refers to how closely synthetic data mirrors the distribution of real data.
- **Utility:** Measures the practical usefulness of synthetic data for tasks such as machine learning or analysis.
- **Privacy:** Ensures that synthetic data does not replicate sensitive aspects of the real data, protecting against privacy risks.
- **Other properties:** Including diversity, coverage, density, and fairness, which assess different facets of the generated data, such as its variety, completeness, and ethical fairness.

Together, these properties provide a comprehensive framework for assessing the quality of synthetic data, ensuring it is suitable for its intended purpose. Over the following sections, different quality properties and their corresponding proposed metrics will be reviewed. Finally, in Table 1 a summary of the reviewed metrics is provided.

### 2.6.1 Fidelity

Fidelity refers to the degree to which synthetic data reflect real data, with higher fidelity indicating greater difficulty in distinguishing between the two datasets. Although fidelity is a key criterion in generative modeling, there is no universally accepted approach to evaluating it. Instead, researchers often rely on a variety of methods that, according to various studies [47, 70], generally fall into three categories: univariate resemblance analysis, multivariate relationship analysis, and data labeling analysis.

Univariate resemblance analysis focuses on preserving the distribution of individual columns in the original data. These measures assess the differences in distribution between corre-

Property	Category	Metric	References
Fidelity	Univariate resemblance analysis	Attribute distributions	[30, 19]
		Dimensional probabilities	[31, 146, 1]
		Kolmogorov-Smirnov test	[10, 19, 34, 160]
		Student t-test	[99]
		Chi-square test	[19]
		Kullback-Leibler divergence	[99, 58, 103]
		Mean absolute error	[123]
		Distance functions	[103, 114]
		Cumulative distribution function	[160]
			Multivariate relationship analysis
Pairwise correlation difference	[58, 103]		
UMAP/PCA	[155, 103]		
Log-cluster metric	[58]		
	Data labeling analysis	Domain experts assessment	[32, 12]
		ML classifiers	[19]
Utility	ML model performance	TSTR	[99, 121, 160, 145]
		Training with augmented datasets	[146]
	ML model metrics	Binary variable: ROC - AUC	[145]
		Continuous var.: RMSE - MAE	[103]
	Classification quality: precision, recall, F1-score.	[145, 103]	
	ML feature importance	SHAP	[55, 145]
Privacy	Identity disclosure	DCR	[115, 103]
		Simulated membership attacks	[31, 115, 103]
	Attribute disclosure	Attribute inference attacks	[31, 103]
	Various privacy risks	Anonymizer	[56]
Others	Data uniqueness	Diversity	[44, 37]
		Rarity Score	[67]
	Synthetic data space	Coverage	[111]
		Density	
	Biases	Fairness	[14]

Table 1: Quality metrics summary.

sponding columns in real and synthetic datasets. Common techniques used in this analysis include statistical tests, distance calculations, and visual comparisons. Classical fidelity assessments, such as comparing attribute distributions between real and synthetic data, are often used for this purpose [30, 19]. Another approach is to compare the dimensional probabilities or dimension-wise occurrences of features between real and synthetic datasets [31, 146, 1].

Some commonly used statistical tests used for the univariate analysis are the Kolmogorov-Smirnov [49] (KS) test, which compare the attribute distributions [10, 19, 34, 160], the Student t-test for comparing the mean values of attributes [99] and the Chi-square test for evaluating the independence of categorical variables [19]. Kullback-Leibler divergence [87] is another well-known metric for assessing how one probability distribution diverges from another, often used for comparing categorical distributions [99, 58, 103].

Additionally, measures such as mean absolute error (MAE) for comparing mean and standard deviation values of each column [123] or distance functions (e.g., Euclidean, cosine, Gower, Hamming distances) also contribute to assessing the univariate fidelity. The cumulative distribution function (CDF), which calculates the probability that a variable is less than or equal to a given value, is also a measure of fidelity. In this method, the CDF of real and synthetic data is compared, and the maximum difference between them serves as a fidelity measure [160].

Multivariate relationship analysis, on the other hand, assesses whether the synthetic data captures higher-order dependencies and interactions among variables. A common technique is the comparison of Pairwise Pearson correlation (PPC) matrices [34, 121, 145]. For example, pairwise correlation difference (PCD) computes the difference between the correlation matrices of real and synthetic data using the Frobenius norm, with smaller values indicating greater similarity [58, 103]. Moreover, dimensionality reduction techniques such as principal component analysis (PCA) or UMAP can be used to visualize and compare the dimensional properties of the two datasets [155, 103]. More advanced techniques, such as the log-cluster metric [150], capture the similarities between the cluster distributions of real and synthetic datasets [58].

Finally, data labeling analysis is an alternative approach where the goal is to classify the datasets as either real or synthetic. Some studies propose qualitative assessments by domain experts, particularly in the healthcare domain, to evaluate the quality of synthetic data [32, 12]. Alternatively, machine learning classifiers can be trained to label the data, providing a more quantitative measure of fidelity by observing how well the classifier distinguishes between original and synthetic data [19].

## 2.6.2 Utility

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. [48] 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 [99, 121, 160, 145]. An alternative approach in-

volves augmenting real training datasets with synthetic samples, as performed in [146], and comparing the performance of models trained solely on real data versus those trained on augmented datasets.

A variety of machine learning models, such as logistic regression, linear regression, random forests, gradient-boosted trees, and deep neural networks, can be utilized for this evaluation. Once the model has been trained, multiple metrics are employed for the purpose of measuring its performance. For example, in the case of a binary target variable (as is the case in classification tasks), the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve is a widely used performance metric. The ratio of ROC-AUC values between models trained on synthetic and real data can be utilized as a quality indicator of the synthetic dataset [103]. In the case of a continuous target variable (as in regression tasks), common performance metrics include the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE).

Furthermore, metrics such as accuracy (the proportion of correctly predicted instances), precision (the proportion of positive identifications that were actually correct), recall (the proportion of actual positives that were identified correctly), and the F1-score (the harmonic mean of precision and recall) are frequently utilized to assess the quality of classification tasks. By comparing these metrics across models trained on real and synthetic datasets, one can assess the utility of the synthetic data.

In addition to performance metrics of the ML models, the analysis of feature importance provides further insight into the quality of synthetic data. By comparing the most salient features identified by models trained on synthetic and real datasets, researchers can ascertain whether the synthetic data effectively captures the same key patterns. A leading method for assessing feature importance is SHAP (SHapley Additive exPlanations), which assigns each feature an importance value for individual predictions [130]. This helps to ensure that synthetic data not only performs well in machine learning models but also reflects the underlying relationships present in the real data [55, 145].

While these methods have been widely applied to tabular data, there are several alternative approaches that have been predominantly developed for synthetic image evaluation but have yet to find their counterparts in tabular data. For instance, metrics such as the inception score (IS) [127] and the Fréchet inception distance (FID) [73] are commonly used for generative image models, offering insights into both the quality and diversity of generated images. However, for tabular data, there remains a gap in the development of analogous metrics tailored to its unique structure and challenges.

Additionally, Sajjadi et al. [126] introduced an innovative method that evaluates the precision and recall of instances produced by GANs. This method considers two distributions,  $P$  (real) and  $Q$  (synthetic), where precision measures the quality of the synthetic samples in  $Q$ , and recall quantifies the coverage of the real data distribution  $P$  by  $Q$ . The combination of these twin metrics forms a two-dimensional score, which allows for a more detailed understanding of potential generative model failures, such as mode collapse, without the need for manual inspection. This approach was later extended by Simon et al. [134], who generalized the precision-recall framework to arbitrary distributions. However, these methods have not yet been applied to tabular datasets, highlighting a gap in evaluation techniques for this data type.

Building on similar concepts, Alaa et al. [3] proposed a three-dimensional metric to evaluate fidelity, diversity, and generalization at the sample level. This approach, which has been tested on tabular, time-series, and image data, introduces  $\alpha$ -Precision,  $\beta$ -Recall, and Authenticity. These metrics offer interpretable probabilistic quantities, enabling a more thorough assessment of how well synthetic data aligns with real-world scenarios across various data modalities.

### 2.6.3 Privacy

Privacy is a critical concern when generating synthetic data, especially in healthcare. While synthetic data offers potential benefits in terms of reducing privacy risks, there is still the possibility that real data records could be exposed or inferred through analysis of the synthetic dataset. Privacy risks include the potential for re-identification of individuals and the leakage of sensitive attributes. The risks generally fall into two main categories: *identity disclosure* [133], where an individual's presence in the dataset is confirmed (membership inference attacks), and *attribute disclosure* [100], where sensitive information about individuals is inferred even without identifying specific records (attribute inference attacks) [18].

Several methods and metrics have been proposed to evaluate the privacy of synthetic data, many of which are based on the evaluation of distance, similarity, or re-identification risk. One widely used method to evaluate whether the records from the real dataset are simply copied into the synthetic dataset, is to compute the distance to the closest record (DCR). The distance, measures the minimal distance between a real record and any synthetic sample. A DCR value of zero indicates a high risk of direct copying [115, 103].

Simulated membership attacks are a common method to assess identity disclosure [31, 115, 103]. For example, in one approach [58], researchers compute the Hamming distance between real and synthetic data samples to evaluate whether synthetic data contains records too similar to the original data. The outcomes are then classified into true positives, false positives, true negatives, and false negatives, and the performance is measured using precision and recall. Another study [160] trains a k-nearest neighbors (kNN) model on synthetic data and evaluates how well the model can predict which real data samples were included in the training set. In this case, a prediction accuracy of 50% would indicate no privacy risk, while higher accuracy suggests potential privacy leakage.

A similar approach is taken to quantify the disclosure risk of sensitive attributes. Different attribute inference attacks have been designed, as in [31, 103], machine learning models are trained on synthetic data to predict sensitive attributes. The accuracy of the model provides a measure of the risk of attribute disclosure

In a more comprehensive approach, *Anonymeter* [56] offers a statistical framework to evaluate various privacy risks, including singling out, linkability, and inference risks. These evaluations are particularly aligned with the European General Data Protection Regulation (GDPR). The framework simulates privacy attacks and evaluates the effectiveness of synthetic data in mitigating these risks. *Anonymeter* also allows for the comparison of synthetic data generated with and without differential privacy techniques (see Section 2.4 for a detailed discussion on differential privacy).

Additionally, differential privacy plays a crucial role in mitigating privacy risks while main-

taining the accuracy of synthetic data. By ensuring that minor changes in the input do not substantially impact the output of a generative model, differential privacy helps prevent re-identification and attribute inference attacks. Beyond being a metric for assessing privacy protection, differential privacy can be integrated into the generative process itself, enhancing the security of synthetic data generation, as reviewed in Section 2.5.

#### 2.6.4 Other properties – Diversity, Coverage, Density, Fairness

In addition to the reviewed properties like fidelity, utility, and privacy, other aspects of synthetic data quality can be assessed to provide a more comprehensive evaluation. Properties such as diversity, coverage, density, and fairness have been explored to measure how well synthetic data reflects different characteristics of real data. However, it is important to note that many of these metrics have been only tested on image data rather than tabular data.

One of the metrics considered is diversity, which evaluates the uniqueness of synthetic data points in relation to the entire synthetic dataset. Diversity measures how similar or distinct each synthetic data point is from others [44]. Models that score high on diversity generate more unique samples, even when the dataset size is large. Various methods have been proposed for measuring diversity. For instance, Donahue et al. [37] use the average Euclidean distance between synthetic data points and their nearest neighbors to assess this property.

Coverage measures the ability of synthetic data to capture the full range of values, especially the extreme points, observed in the original data. This metric focuses on how well the generative model replicates the minimum and maximum values of the features present in the real dataset. Closely related to these is the density metric, which goes beyond simple precision. While precision only checks whether synthetic samples fall within the real data manifold, density measures how many real data neighborhoods (defined by k-nearest neighbors) contain synthetic data samples [111]. This metric enhances robustness to outliers and can provide a more accurate assessment of how well the synthetic data populates the same space as real data.

Both density and coverage were introduced to address shortcomings in the traditional metrics of precision and recall. Specifically, precision and recall tend to be sensitive to outliers and computationally inefficient when applied to high-dimensional data. The coverage metric, in particular, improves upon recall by measuring the proportion of real samples that have at least one synthetic sample within their neighborhood. This method better captures the diversity and extent of the real data distribution, but like many other advanced metrics, it has primarily been validated on image datasets.

Another similar proposal is the rarity score [67], which measures the individual rarity of each synthetic sample by estimating the local density around it on the real data manifold. Unlike other metrics that focus on sets of images, the rarity score evaluates each synthetic sample individually, providing insights into how well rare or underrepresented samples are generated. However, like other metrics, this too has been developed primarily for image data.

Lastly, the concept of fairness in synthetic data generation is gaining attention. It refers to the requirement that the artificially created data preserves equitable representation and treatment of all sensitive demographic groups present in the original dataset. In practical terms, this means the generative process must ensure that protected attributes do not become corre-

lated with unfavorable outcomes in the synthetic data, either through disproportionate representation or through the perpetuation and amplification of existing real-world biases.

Studies have shown that many GAN-based approaches, rather than eliminating bias, exacerbate it in the synthetic output [63]. To address this, initial solutions like FairGAN [153] and DECAF [142] have been proposed, incorporating fairness constraints into the generative process to ensure that synthetic data is more equitable.

To evaluate how well a model deals with equity and representation issues, different metrics have been introduced to measure fairness in synthetic data. For instance, the *log disparate impact equity metric*, proposed by Bhanot et al. [14], compares the proportions of different subgroups in the synthetic data to those in the real data. This metric checks whether the proportions of these subgroups are preserved in the generated data, providing a more detailed analysis at the subgroup level. While a method may succeed in reflecting fairness across the dataset overall, this approach ensures that subgroup-level fairness is also considered.

## 2.7 Multimodal data

Multimodal data refers to datasets that encompass multiple types of information captured from various sources and formats. These modalities can include tabular data, images, text, audio, and more. In the healthcare context, multimodal data might comprise electronic health records (EHRs), disease registries, vital signs, medical imaging such as MRI, physician-authored clinical notes, blood test results, etc. The integration of these diverse data types provides a comprehensive view of patient health, enabling more accurate analysis and decision-making.

The importance of multimodal data is increasingly recognized in healthcare, where a multimodal, data-driven approach is driving advancements in smart healthcare systems [22, 131, 106]. These systems utilize multiple data sources for applications ranging from disease analysis to triage, diagnosis, and treatment. The integration of various data types enhances the precision and personalization of healthcare interventions, making it a key enabler for innovations such as computer-aided diagnosis and treatment recommendation systems.

Just like with single-type data, synthetic multimodal data generation has become essential. Synthetic multimodal data is used to address the same issues of regular synthetic data: data scarcity, privacy concerns, and class imbalances. However, generating multimodal data poses unique challenges [91], such as multimodal representation and alignment. Multimodal representation refers to learning a unified representation of data from diverse sources, while multimodal alignment involves identifying relationships between components from different modalities, which is crucial for decision-making models. The heterogeneity of statistical properties in different modalities makes both tasks difficult [62, 81].

Several recent advances have been made in generating synthetic multimodal data, although challenges remain. For instance, the complexity of combining data with varying structures and distributions—such as text, images, and tabular data—requires innovative approaches to align and fuse these modalities effectively [39, 17, 161].

One notable example is MVAESynth [93], a framework based on a multimodal variational auto-encoder (MVAE), which was developed to generate realistic synthetic data. Though primarily applied to generating synthetic social network profiles, this model provides a foun-

dition for further development in other domains, including healthcare. Another approach, proposed by Li et al. [91], uses a deep multimodal generative adversarial network (DMGAN) to generate synthetic multimodal data for classification tasks, although the model was tested on non-healthcare datasets.

In the healthcare domain specifically, Haleem et al [65]. propose a novel approach to generate synthetic multimodal data from real-time electronic health records and physical records. This method addresses the need for generating high-quality, privacy-preserving synthetic healthcare data while considering both medical and physical attributes.

## 2.8 Integration in federated learning

Federated Learning (FL) [102] is a decentralized machine learning framework that enables multiple parties to collaboratively train a global model while keeping their local datasets private. Rather than sharing raw data, FL involves training local models on each client's data and exchanging only model parameters, such as weights or gradients, to aggregate them into a global model. This approach was initially developed to address privacy concerns and increase data accessibility in scenarios where data transfer to a centralized server is not feasible due to privacy, legal, or logistical constraints[76]. As a result, FL has emerged as an effective solution for collaboratively training machine learning models without the need to exchange raw data.

Despite its promise of enhanced privacy, FL faces several challenges. Chief among these is the risk of privacy breaches, as well as fairness concerns regarding potential biases that may arise in favor of or against certain clients or demographic groups [118]. For instance, a malicious server could infer sensitive information from the updates shared by clients or tamper with the training process. Similarly, an adversarial client could deduce confidential information from other participants updates or manipulate the parameter aggregation process to corrupt the global model [94]. Studies have demonstrated that FL systems can be vulnerable to various forms of privacy attacks, particularly during the phase when model weights are exchanged between clients and servers [95, 136, 151].

Federated learning can be categorized into three types, based on the distribution of data features and sample spaces among users [76]:

- **Horizontal Federated Learning** occurs when data across different users have overlapping features but distinct samples [158]. This is applicable when different users collect similar types of data, but from different populations. This is the initially considered case in the FLUTE project.
- **Vertical Federated Learning** applies when users share the same sample space (i.e., the same individuals or entities) but collect different types of features for each sample [148]. In this case, the data is divided vertically based on features.
- **Federated Transfer Learning** is used when both users and features have little overlap across datasets [157]. In these cases, transfer learning techniques can be applied to enhance the performance of models by transferring knowledge from one domain to another.

The use of federated learning has gained considerable traction in various domains, particularly in healthcare, where privacy and data security are critical concerns. The sensitive nature of healthcare data, often regulated by strict privacy laws, makes FL an attractive solution for enabling collaborative learning across institutions without sharing raw data [120]. For instance, a recent study [88] proposes a digital healthcare framework for patients with disabilities using federated deep learning schemes, employing federated deep convolutional neural networks (FL-DCNNs). Similarly, another work [15] introduces a federated AI architecture for stroke prediction, utilizing artificial neural networks (ANNs) on real-world stroke cases. This architecture can be implemented in healthcare wearable devices for real-time, accurate predictions, making it both computationally efficient and effective.

As synthetic data generation also aims to protect privacy while expanding data availability, its integration with federated learning is a natural progression. Synthetic data generation can augment federated models, particularly when there is insufficient data available at local nodes. In the same way, generating synthetic data in a federated environment can facilitate the access to more original data without privacy concerns. Recent studies have explored combining FL with synthetic data generation. For example, the Federated Tabular Diffusion model (FedTabDiff) [129] leverages diffusion models to generate high-quality tabular data in a decentralized manner, enabling entities to train a generative model collaboratively without compromising data privacy. In another study [124], a federated learning framework called FedCSCD-GAN is proposed for cancer diagnosis. This framework incorporates GAN-generated synthetic data to enhance the classifier generalization, stabilize training, and reduce overfitting, particularly when dealing with limited labeled samples.

## 3 Datasets

According to the Project Objective PO3, synthetic data generator models should be developed and synthetic databases generated using these generator models. Statistical tests should validate these approaches.

The proposed methodology will be designed to be applicable to several topics within the health domain. The focus is specifically on prostate cancer; however, other databases are considered to check the robustness of our algorithms.

### 3.1 Prostate cancer datasets

Two public data sets are utilized: the Prostate Imaging and Cancer Analysis Information (PI-CAI) dataset, available at <https://pi-cai.grand-challenge.org/DATA/> (accessed on 17 April 2025) and the Cancer Imaging Archive (CIA) dataset available at <https://www.cancerimagingarchive.net/collection/prostate-mri-us-biopsy/> (accessed on 17 April 2025).

The PI-CAI data set comprises 1500 samples, each containing 12 features related to prostate cancer. The attribute `case_csPCA` is the primary outcome variable, indicating the presence of clinically significant prostate cancer. Certain features, such as `patient_id`, `study_id`, and `mri_date`, are excluded from this analysis due to their irrelevance for predicting `case_csPCA`.

and the potential privacy concerns they pose. Additionally, the attribute `histopath_type`, which refers to the procedure used to sample lesion tissue, is disregarded in this study because it does not directly impact the diagnosis of prostate cancer.

The attribute `lesion_GS` represents the Gleason Score (GS), a grading system that evaluates the aggressiveness of prostate cancer based on the microscopic appearance of cancer cells. Another important feature, `lesion_ISUP`, simplifies the Gleason Score into five distinct ISUP grades to better predict clinical outcomes. Furthermore, `case_ISUP` denotes the highest ISUP grade among the sampled lesions. Additional features include `psa`, which represents the prostate-specific antigen level, and `psad`, which is the prostate-specific antigen density. In cases where `psad` is missing, it can be computed by dividing PSA by the prostate volume (`psad_computed`).

Since `case_ISUP` reflects the highest risk grade in a sample, initially it is considered instead of `lesion_ISUP`. However, a correlation analysis revealed a substantial correlation coefficient (0.87) between `case_ISUP` and `case_csPCA`, leading to the exclusion of `case_ISUP` as a predictor. Consequently, `patient_age`, `prostate_volume`, and either `psa`, `psad`, or `psad_computed` are selected as predictors for the output `case_csPCA`.

The CIA data set, also a public repository, provides a wide range of information on prostate cancer. However, unlike the PI-CAI data set, the `case_csPCA` attribute is not available. Despite this limitation, the data set contains `psa` and `prostate_volume` values for each of its 24,783 samples. The attribute `patient_age`, however, is missing. Given the size and diversity of the CIA data set, it offers a valuable resource for exploring correlations and patterns in prostate cancer. The objective is to generate synthetic values for `patient_age` and assign synthetic `case_csPCA` values to each sample based on the available information.

## 3.2 Breast cancer datasets

Our study utilizes two publicly available breast cancer datasets. The first dataset, referred to as BC-MLR, is the Breast Cancer dataset from the UCI (University of Chicago, Illinois) Machine Learning Repository, available at <https://archive.ics.uci.edu/dataset/14/breast+cancer> (accessed on 17 April 2025). The second dataset, BCNB, is the Early Breast Cancer Core-Needle Biopsy WSI dataset, accessible at <https://paperswithcode.com/dataset/bcdalnmp> (accessed on 17 April 2025).

The BC-MLR data set comprises 286 instances with nine attributes related to patient characteristics and tumor details. Patient-specific data include age at the time of diagnosis, represented by the `age` attribute, as well as the patient menopausal status at diagnosis, noted in the `menopause` feature. Tumor-related information is captured through variables such as `tumor_size`, which records the maximum diameter of the excised tumor, and `deg-malig`, reflecting the tumor degree of malignancy and aggressiveness.

Additional features in the BC-MLR data set address the extent of cancer spread, with `inv-nodes` indicating the number of lymph nodes involved in metastatic breast cancer, and `node-caps` providing information on whether the tumor has replaced the lymph and penetrated the capsule. Moreover, the data set records the affected breast side (`breast`) and the specific breast quadrant where the tumor is located (`breast-quad`). The data set also includes `irradiat`,

which specifies whether the patient has undergone radiation therapy. The target variable in this data set is `Class`, which represents whether there is cancer recurrence after treatment.

On the other hand, the BCNB data set includes information from 1058 breast cancer patients. The `age` attribute records the patient age at the time of diagnosis, while tumor characteristics such as `tumor size` and `tumor type` provide detailed descriptions of the tumor. Additionally, the data set captures molecular markers such as the expression of estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2), represented by the features `ER`, `PR`, and `HER2`, respectively. The data set also includes `HER2 expression`, which provides the results of an immunohistochemistry (IHC) test, used to predict how well the cancer will respond to treatment.

The BCNB data set further includes the `histological grading` feature, which assesses how closely the tumor cells resemble normal breast cells. Information regarding surgical interventions, specifically the type of surgery performed to remove nearby lymph nodes, is contained in the `surgical` feature. Other features of note include the `Ki67` score, which measures the rate at which cancer cells are dividing, and the `molecular subtype`, a classification system grouping breast cancers based on shared characteristics and hormone receptor status. Finally, the data set provides information on lymph node involvement through features such as the `number of lymph node metastases` and `ALN status`, indicating whether the lymph nodes in the underarm area are affected.

Since ideally the database would have the same features, we focus on common features between the data sets. Therefore, particular attention is given to the `age` attribute, `tumor size`, and `number of lymph node metastases`, as these variables are present in both the BC-MLR and BCNB data sets.

### 3.3 Cardiovascular disease datasets

Two datasets are used in the field of cardiovascular disease for the second algorithm. 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) Data Set. This dataset is from Liverpool John Moore's University available at <https://iee-dataport.org/open-access/heart-disease-dataset-comprehensive> (accessed on 17 April 2025). The `fiv-CardioDB` dataset contains 1190 instances with 11 features.

The second dataset, known as `Ind-CardioDB`, was developed by Doppala et al. at Lincoln University College. It is available at <https://data.mendeley.com/datasets/dzz48mvjht/1> (accessed on 17 April 2025) The `Ind-CardioDB` 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* (`restingrelectro`) results, 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.

## 4 UMAP-PSDG Algorithm

Checking data from the clinical partners, the common case of lack of information appeared. This is due to technical issues, transcript errors, or differences between descriptors considered in different health centres, leading to the need for data imputation and partial data generation techniques.

A first algorithm introduces a novel methodology for partially synthetic tabular data generation, designed to reduce the reliance on sensor measurements and ensure secure data exchange. Using the UMAP (Uniform Manifold Approximation and Projection) visualization algorithm to transform the original, high-dimensional reference data set into a reduced-dimensional space, we generate and validate synthetic values for incomplete data sets. This approach mitigates the need for extensive sensor readings while addressing data privacy concerns by generating realistic synthetic samples. The proposed method is validated on prostate and breast cancer data sets, showing its effectiveness in completing and augmenting incomplete data sets using fully available references. Furthermore, our results demonstrate superior performance in comparison to state-of-the-art imputation techniques. This work makes a dual contribution by not only proposing an innovative method for synthetic data generation, but also studying and establishing a formal framework to understand and solve synthetic data generation and imputation problems in sensor-driven environments.

This section presents the main results from our work in this first algorithm. A more complete study can be found in our publication [97]:

Lázaro, C., & Angulo, C. (2024). Using UMAP for Partially Synthetic Healthcare Tabular Data Generation and Validation. *Sensors*, 24(23), 7843. <https://doi.org/10.3390/s24237843>

### 4.1 Motivation

In [79], synthetic data are classified into two principal categories: fully synthetic and partially synthetic. Fully synthetic data sets are entirely artificial and contain no original data. In contrast, partially synthetic data sets replace only certain attributes with synthetic values, resulting in a hybrid data set that combines real and synthetic data. This approach is particularly

advantageous for handling sensitive attributes, where synthetic versions can be employed to substitute sensitive values while maintaining the utility of the non-sensitive data.

In cases where the challenge is not related to sensitive data but rather missing values, the focus shifts to data imputation [137]. Given that imputed values are also synthetically generated values, there exist significant overlaps between the research domains of data imputation and synthetic data generation. Specifically, the outcome of a data imputation process essentially results in a partially synthetic data set. However, the expression “data imputation” is typically applied in the context of addressing missing data, whereas “synthetic” or “partially synthetic data” are commonly employed either to mitigate concerns about sensitive data or a data augmentation procedure.

Our algorithm focuses on the generation of tabular data in the context of healthcare, introducing a novel methodology that leverages **data visualization tools** to enhance the generation of synthetic data. To the best of our knowledge, no existing methods employ visualization tools to analyze real-world reference data structures for the generation of partially synthetic data.

Hence, this work establishes a totally new research direction in partially synthetic data generation. We aim to provide a foundation for the use of dimensionality reduction tools for data synthesis purposes. Ultimately, this line of research aims to extend to visualization-based fully synthetic data and provide a baseline for comparing this approach with more established generative models, including VAEs, GANs, and diffusion models.

#### 4.1.1 Fully and partially synthetic data

Synthetic data generation involves either partial synthesis, where only some original records are replaced, or full synthesis, where the entire data set is substituted with a synthetic data set of the same size and composition [38].

The distinction between fully and partially synthetic data leads to the differentiation between fully and partially synthetic methods. Fully synthetic methods aim to replicate the statistical properties and relationships inherent in the original data without retaining any direct links to real individuals. While this approach provides robust privacy protection, it may compromise the data fidelity. Conversely, partially synthetic methods seek to combine real and artificial data [110]. These data sets are typically used when only certain attributes are sensitive, allowing non-sensitive parts to remain unchanged. This approach seeks to balance privacy and data utility.

#### 4.1.2 Data visualization and dimensionality reduction tools

The field of data visualization encompasses a range of techniques and tools that facilitate the comprehension of the structural organization and distributional characteristics of complex data sets. The visualization of high-dimensional data presents a significant challenge. To address this issue, data visualization techniques have been developed to reduce data dimensionality from highly dimensional space to a 2D or 3D representation. The most commonly utilized tools are t-distributed stochastic neighbor embedding (t-SNE) [143], Uniform Manifold Approximation and Projection (UMAP) [101], and TriMap [5]. Each of these methods possesses distinctive strengths and limitations, particularly in regard to the nature of the data being processed.

Data visualization tools have demonstrated their utility in various health-related studies, particularly with UMAP. For instance, UMAP is used to visualize electronic health records (EHRs) [77], to extract patient features for detecting depression [105], or identify prognostic factors [149] in leukemia. On the other side, dimensionality reduction has also demonstrated significant value for imputation purposes [4].

In the context of this methodology, reproducibility and effective cluster visualization are of paramount importance. Specifically, we are interested in distinguishing clinically significant prostate cancer samples within the PI-CAI data set. All three methods were applied to visualize the most comprehensive data set available for the two types of cancer, PI-CAI and BC-MLR, after normalizing the data to ensure consistency in the results. It can be checked in [97] that UMAP offers a more distinct separation of classes compared to the other methods, highlighting UMAP's superior ability to organize the data meaningfully.

## 4.2 Partially synthetic methodology proposal

The proposed procedure is designed to facilitate the exchange of medical data, with the primary goal of augmenting or completing data sets for use in data-driven solutions. Let us imagine hospital 1 with a private, complete, and small data set. On the other side there is hospital 2, with a larger data set, but in this case uncompleted. Our methodology allows hospital 2 to leverage the complete information in hospital 1 without sharing the actual data.

The methodology begins by extracting information from the complete data set in hospital 1, which is named the reference data, and transforming it into a two-dimensional space using UMAP. In a second step, using this transformed data map, along with the trained UMAP function, synthetic values are assigned to the incomplete data set (from hospital 2). These newly assigned synthetic values are then mapped back into the 2D representational space to evaluate their congruence with the reference data.

The proposed methodology is a general procedure able to generate artificial data which, as previously discussed, can be categorized into fully synthetic, partially synthetic, and imputed data. These different results are achieved through different methods, as illustrated in Figure 6. There, it is highlighted how fully synthetic methods generate data sets that contain no real data from the reference or incomplete data sets. In contrast, partially synthetic methods produce data sets that combine real data from an incomplete data set with artificial data generated from the available information. Finally, imputation methods, which typically address missing data, result in data sets that contain a mix of real and artificial data. These imputation methods can either generate the missing values from only the actual incomplete data set (option A) or complement it with a reference data set containing real-world data (option B).

Since partially synthetic data aim to protect sensitive information and imputed data seeks to resolve missing data issues, the artificial data in the resulting data sets usually exhibit different distributions. Partially synthetic data typically include some synthetic features that are consistent across all entries. Conversely, the artificial values in an imputed data set do not necessarily correspond to the same feature for all entries.

The proposed methodology aims to integrate both real-world reference data and incomplete data to generate a complete partially synthetic data set. This approach leverages a structured framework to ensure both data utility and privacy preservation. The overall workflow of the

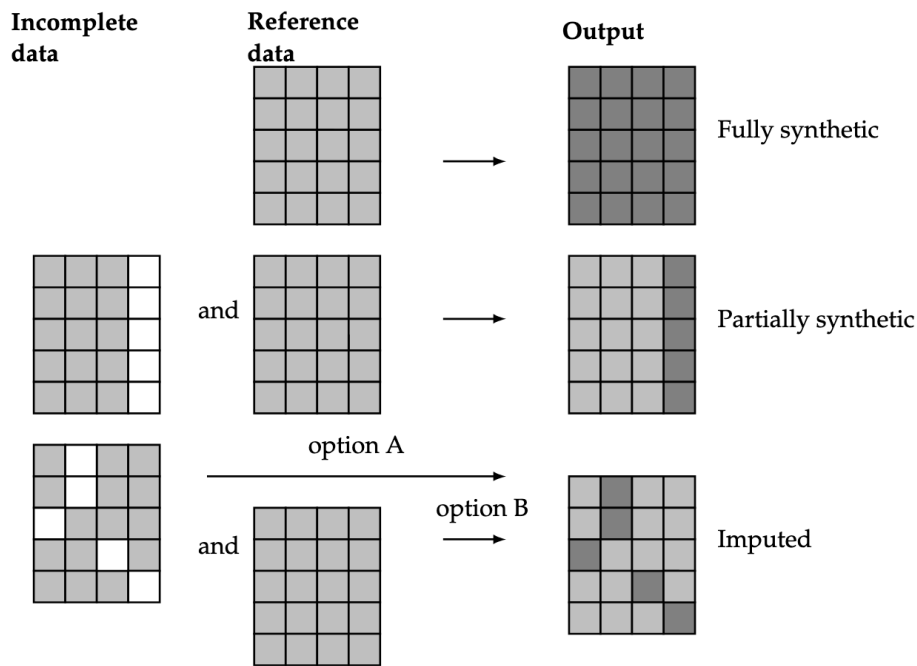


Figure 6: Representation of real (light gray) and artificial (dark gray) values for missing data (white) on fully synthetic, partially synthetic, and imputed data sets.

methodology is illustrated in Figure 7, providing a visual representation of the process steps and their interconnections.

To operationalize this framework, we have developed an algorithm that outlines the detailed steps required to execute the methodology, including preprocessing, synthetic data generation, validation, and final data set compilation. In this section, the different steps presented in the algorithm are described.

As previously described, the initial setup consists of a reference database and an incomplete database, which contains the same features as the reference data, except for the missing feature.

**Step 1: Preprocessing and setup.** The first step, before applying the methodology, is to ensure uniformity in data structures between the incomplete and reference databases. This setup involves aligning units, sampling groups, data collection methodologies, and normalization techniques. Let us assume that the range  $M$  of finite possible values exist for the feature  $x$  in the reference database.

**Step 2: Synthetic data generation.** Using the  $M$  possible values of the missing feature,  $M$  different entries can be created for each sample by assigning all different values of the range. Consequently, a resulting generated database is obtained. Considering that the incomplete database contains  $N_i$  samples, the generated database contains  $N_i \cdot M$  samples and  $d$  features ( $d - 1$  of these are real-world features and one feature is synthetically generated).

**Step 3: UMAP transformation.** The core of the proposed methodology lies in utilizing the visualization tool UMAP to transform data from  $d$  dimensions to coordinates in two dimensions. By performing these transformations, the UMAP coordinates of the generated data can

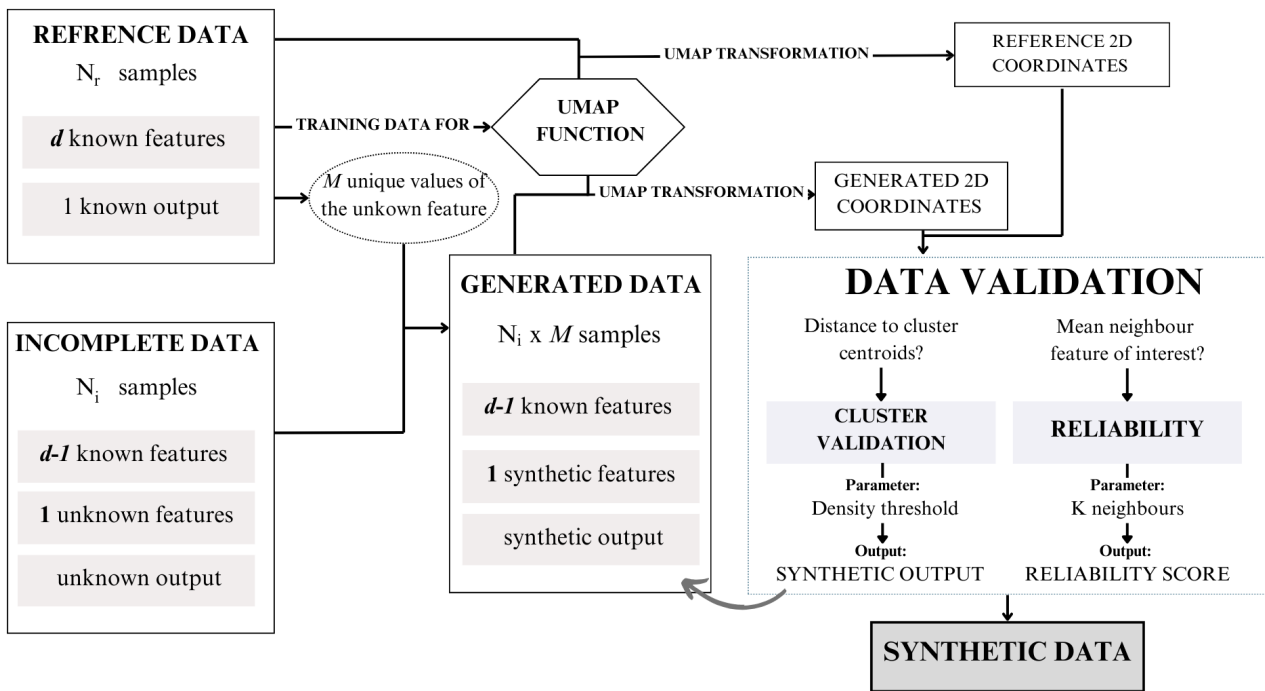


Figure 7: Synthetic data generation framework proposal.

be obtained and compared to the coordinates of the reference data.

**Step 4: Reliability score assignment.** Now, a validation step, defined as the *reliability score assignment* determines which of the generated samples are suitable and which do not align well with the reference data. If the reliability score  $r_c$  of a certain sample exceeds a threshold, then the generated  $i$ -th sample is validated. For a single original sample in the incomplete database, there may be multiple suitable values for the missing feature, resulting in the generation of additional samples. Therefore, the methodology not only imputes the real missing value, but also creates additional suitable samples, which constitute synthetic data.

**Step 5: Cluster validation (if applicable).** For the last step, we can consider that the output is also unknown. Since the output is known for the reference data, instances can be grouped into clusters, so the output of the  $N_i$  entries can be assigned based on the distance between each generated coordinate and the cluster's centroid coordinates.

Setting a minimum reliability score threshold for validated samples ensures that only high-quality synthetic data points are retained for downstream applications. This threshold serves as a filter to enhance the reliability and usefulness of the generated data.

**Step 6: Output.** Upon completing the reliability score analysis and, if applicable, the cluster validation process, the algorithm produces a final set of synthetic samples. These samples are selected based on meeting the reliability score threshold, ensuring their consistency with the reference data. Additionally, for scenarios where the target variable is unknown, the algorithm assigns a predicted target value through the cluster validation process. The resulting data set consists of high-quality synthetic samples that are suitable for downstream applications.

## 4.3 Results

In this section, the proposed methodology is applied to generate synthetic data using the data sets introduced in Section 3. Since the true values for the missing data in the incomplete data sets are not available, direct fidelity validation of the generated synthetic data cannot be performed.

The experiments considered for algorithms validation are the following:

1. Prostate Cancer use case. CIA database is the reference data and PI-CAI database is the database to be completed with synthetic data, the `patient_age` feature.
2. Breast Cancer use case. BC-MLR is the reference data and BCNB is the incomplete data set. In this case, we consider to replace original data with the iterative generation of three synthetic features. For easiness, features being synthetically generated are those having highest correlation among both databases: `deg-malig`, `node-caps`, and `irradiat`.

### 4.3.1 CIA synthetic data generation from PI-CAI

In our experimental setup, we aim to generate synthetic samples for the CIA database using the PI-CAI data as a reference. Therefore, the reference data set comprises three features and a target variable per sample and the incomplete data set lacks one feature and the target variable.

After applying normalization and UMAP transformation to the reference data set, we ascertain the 2D coordinates of the PI-CAI reference samples. Aside from some outliers, the output variable facilitates the definition of clusters.

The `patient_age` feature in the reference database (PI-CAI) shows values from 35 to 92, therefore the number of possible values for the feature of interest is  $M = 58$ . By assigning to each sample on the incomplete database all the possible ages, the resulting synthetic data set is composed of  $58 \times 24,728 = 1,434,224$  potential entries.

The evaluation and validation of the generated data reveal several key findings. Cluster validation depends on the density threshold utilized to define the cluster area. A marked reduction in the number of validated samples, particularly within the cluster `case_csPCA= 0`, is evident when the density threshold exceeds the value of 6. This significant decrease is also reflected in the total percentage of validated samples, which notably drops at this threshold value. This reduction indicates that for density thresholds greater than 6, the area around the cluster `case_csPCA= 0` centroid fails to encompass the entire cluster. For the cluster `case_csPCA= 1`, this phenomenon occurs at a density threshold of 7. Based on the findings, the density threshold for this reference database is set to 6.

### 4.3.2 BCNB synthetic data generation from BC-MLR

To evaluate the reproducibility of the proposed methodology, the same procedure described in the previous section was applied to breast cancer data. In this experiment, the reference data set is the Breast Cancer Machine Learning Repository (BC-MLR) while the incomplete data set is the Breast Cancer Core-Needle Biopsy (BCNB) data set. As presented in Section 3,

both data sets share three common features: patient age, tumor size, and the number of lymph nodes containing metastatic cancer. Given the relatively small size of both data sets, the goal is not only to generate synthetic features, but also to augment the number of samples. In addition, since the BCNB data set is missing the output variable, the intention is also to obtain this target value.

Unlike in Section 4.3.1, where a specific variable was generated, the goal here is to generate several synthetic features for the BCNB data set. The only condition to generate a new feature and add it to the known features in BCNB is that they must be present in the BC-MLR. The correlation between the BC-MLR variables and the target feature `class` is considered to determine which features to generate. Table 2 shows the variables from BC-MLR and their corresponding correlations with the target attribute. The correlation between the known features is written in gray. Three other features with significant correlations with the target variable are chosen to be generated: `node-caps`, `deg-malig`, and `irradiat`.

Table 2: Feature correlation with target variable.

Feature	Correlation
age	-0.072
menopause	0.052
tumor-size	0.13
inv-nodes	0.27
node-caps	<b>0.24</b>
deg-malig	<b>0.3</b>
breast	-0.059
breast-quad	0.037
irradiat	<b>0.19</b>

The procedure followed is the same as described in Section 4.3.1, but in this case it is applied iteratively, generating one feature at a time. In the first iteration, the reference data set consists of the three known attributes, and it is projected into the UMAP space. In subsequent iterations, the attribute to be generated is added to the reference data set, resulting in a new UMAP representation for each iteration. Each time the reference data change, the density threshold to define the cluster area should be adjusted; in this case, for all iterations the density threshold was set to 1. This iterative approach allows the generation of multiple features, each producing a distinct UMAP projection based on the updated reference data.

In the data generation process, the reliability score threshold is set to  $r_c = 1$  for each iteration. The first feature to be generated is `deg-malig`, which can take a value of 1, 2, or 3, resulting in a total of  $3 \times 1058 = 3174$  possible synthetic samples. Of these, 2558 samples were validated, accounting for 80.59% of the possible combinations.

The second generated feature is `irradiat`, which can take a binary value ('yes' or 'no'). Based on the previously generated data set containing 2558 samples, this step produces a

data set of 5092 samples. Finally, the feature `node-caps` is generated, resulting in a final data set of 9427 samples.

By applying the proposed methodology iteratively, the initial incomplete data set with 1058 samples and three attributes is transformed into a partially synthetic data set with 8.9 times more samples (9427), six attributes, and the synthetic target feature (`Class`).

### 4.3.3 Discussion of results

The proposed methodology has demonstrated a robust capability to generate synthetic data, achieving an augmentation of between thirteen and thirty times—depending on the selected reliability threshold—the amount of incomplete data in the context of prostate cancer databases. In the case of breast cancer, the original data sets are smaller, yet the methodology successfully yields a partially synthetic data set that is approximately nine times larger than the original. Notably, this augmentation includes the generation of three attributes along with the target variable.

However, the proposal relies on a reference data set, which poses a potential limitation. In practical use cases, especially in clinical settings, the availability of a well-structured reference data set may not always be guaranteed. If such a data set is absent, the performance of the methodology could be compromised. To address this issue, alternative approaches such as using publicly available databases or synthetic reference data sets could be explored.

One key advantage of generating this large volume of data is that it allows for machine learning models' generalization capabilities to be improved in low-data environments. The ability to generate high-quality synthetic data may reduce overfitting in models and allow for more robust decision making in clinical and other sensitive domains.

Yet, the advantage and disadvantage of generating a large amount of data is that, without knowing the expected value, it becomes challenging to differentiate between imputed and synthetic samples. This ambiguity benefits synthetic data generation, as all the data can be utilized without privacy concerns, given that the small percentage of real data cannot be distinguished.

However, the developed algorithm may not be suitable for simple imputation, that is, to replace missing values with a single estimated value (like the mean, median, or a predicted value from a model), treating the imputed value as certain, as the correctly imputed samples are mixed with synthetic data. Despite this, since multiple values are generated for each entry, the method could be adapted for multiple imputation, that is, to generate several plausible versions of the complete dataset (typically 5-20), each with different imputed values drawn from a probability distribution. Hence, each generated value could be assigned to the missing data in different data sets, creating multiple imputed data sets for analysis. Given that the number of assigned values for each entry is not constant, adjustments should be made to the number of iterations for multiple imputation.

While this study applied UMAP for dimensionality reduction on a data set with relatively low dimensionality, reducing a maximum of six dimensions to two, the full potential of UMAP in handling very high dimensional data remains to be explored. UMAP is particularly powerful in applications involving hundreds of dimensions, such as genomic or proteomic data sets.

In summary, while the methodology offers clear advantages in terms of synthetic data generation and augmentation, its dependency on reference data and its potential limitations in simple imputation tasks warrant further investigation. Nonetheless, its broad applicability in fields with available reference data sets makes it a valuable tool for data augmentation, particularly in healthcare.

## 5 Iterative UMAP-FSDG

Building on the previously developed partially synthetic data generation algorithm utilizing data visualization techniques, UMAP-PSDG, 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). This approach has been successfully applied to three healthcare domains: prostate cancer, breast cancer, and cardiovascular disease. In short, results show that this method represents a robust solution for generating secure, high-quality synthetic healthcare data, effectively addressing data scarcity challenges.

This section presents the main results from our work in this iterative version for fully synthetic tabular data generation. A more complete study can be found in our publication [96]:

Lázaro, C., & Angulo, C. (2024). Iterative Application of UMAP-Based Algorithms for Fully Synthetic Healthcare Tabular Data Generation. *Algorithms*, 17(12), 591.  
<https://doi.org/10.3390/a17120591>

### 5.1 Motivation

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. 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.

Building on the concept of partially synthetic data, our previous study proposed algorithms that use visualization tools, in particular, UMAP, to validate and enhance synthetic data generation [97]. 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 which 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 to that of synthetic samples. However, while partially synthetic data reduces certain privacy risks, it 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 the algorithm development 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. Main concern with iterative processes is about degradation. In the same form that happens with error propagation in iterative methods, it should be checked how synthetic data can degrade when it is iterated mixing real-world with synthetic generated data.

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. Finally, only samples that exceed a certain reliability score threshold are included in the resulting *partially synthetic dataset*.

## 5.2 Iterative UMAP FSDG methodology

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, which serves as the incomplete dataset. For fully synthetic data generation, however, the process involves an iterative feature generation approach, as detailed in Figure 8:

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.

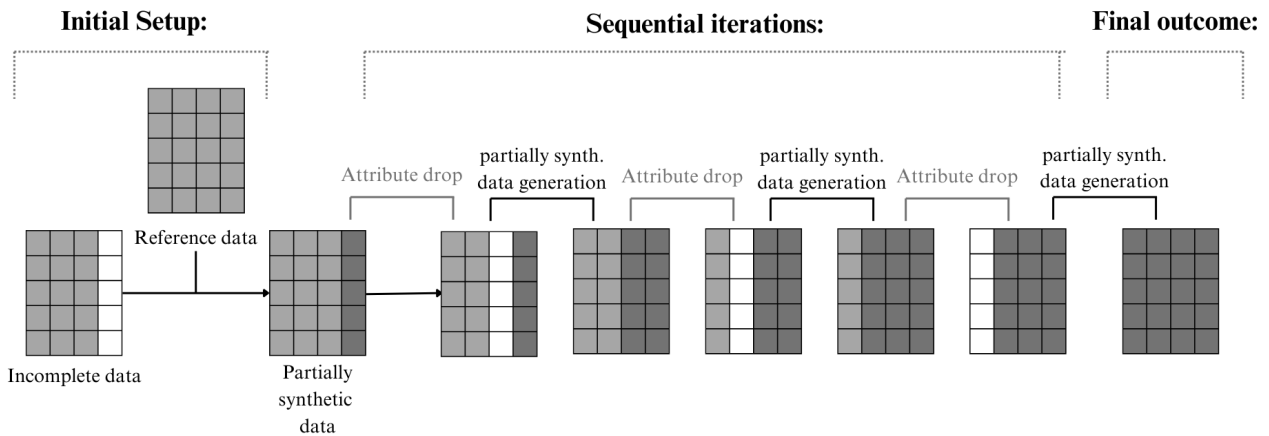


Figure 8: Algorithm workflow

In the following, the algorithm is described in detail.

**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.

**Dataset partitioning and setup** The initial stage of the methodology involves data preparation, tailored to the nature of the available data. 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 data set helps create a more robust synthetic dataset that accurately reflects the diversity of the original data.

**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. Quantization is a key element for continuous features to allow discrete, controlled synthetic values for comparison and validation.

**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. In this case, the interest is to define a cluster area that comprises a wide majority of the samples, therefore the threshold is set to 90%.
- $\epsilon$  for reliability score assignment: The reliability score quantifies the validity of each synthetic sample based on feature disparity.

## 5.3 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.

### 5.3.1 Applying the algorithm

**UMAP transformation and feature distribution** The initial phase of the methodology involves transforming the available data into a lower-dimensional space using UMAP. This transformation provides a two-dimensional representation, enabling the visualization of feature distributions within the datasets. These visualizations are crucial for validating the quality of the generated synthetic data.

In the case of the cardiovascular data (Figure 9), an additional level of differentiation is observed within the primary clusters. Further analysis reveals that this secondary separation corresponds to the feature `fastingbloodsugar`, highlighting UMAP's ability to capture subtle structures within the data.

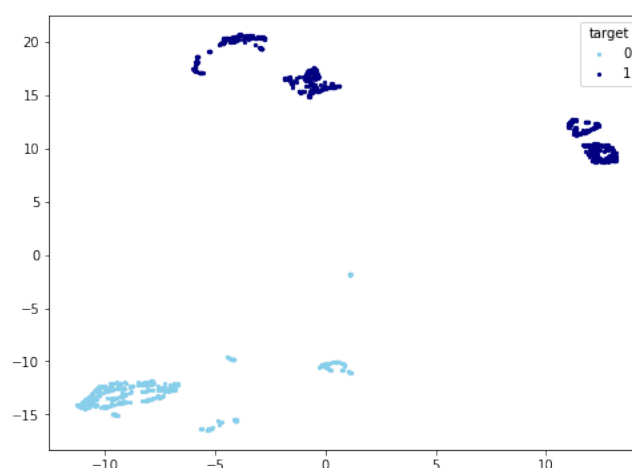


Figure 9: Cardiovascular diseases data visualization (Ind-CardioDB).

**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 1,500 samples with 425 samples representing `case_csPCA = 1` and 1,075 representing `case_csPCA = 0`. While the dataset is imbalanced, the `case_csPCA = 1` class is sufficiently represented within the UMAP-transformed space. The data is divided into a reference dataset (500 samples) and an incomplete dataset (1,000 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, 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 will be 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).

**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. 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).

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 data set 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.

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

**Hyperparameter tuning for synthetic data validation** Hyperparameters, such as the density threshold for defining cluster areas and the  $\epsilon$  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. In order to determine the optimal value, the percentage of cluster validated samples against the different density thresholds is considered. 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.

The  $\epsilon$  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.

To calculate  $\epsilon$ , 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.

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  $\epsilon$  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  $\epsilon$  values for each case.

Table 3: Mean feature disparity for prostate cancer reference data (PI-CAI)

Feature	Mean $f_{disp}$	$\varepsilon$
patient_age	1.18	1
PSA	3.77	4
prostate volume	5.67	6

### 5.3.2 Discussion of 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 involves 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 1,000 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 1,058 samples with three attributes. In this case, since part of the input dataset did not contain all the attributes we aim to generate, the partially synthetic data generation method [97] was applied to generate the missing attributes. This resulted in a partially synthetic dataset of 9,427 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 1,190 samples. By iteratively generating synthetic values for all attributes, the algorithm produced a fully synthetic dataset of 50,476 samples.

## 6 Validation

The evaluation or validation 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 receives less attention compared to synthetic image data. For example, [72] 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 is insufficient due to privacy restrictions.

The fidelity of synthetic data is a measure of how well the synthetic data represents the statistical properties of real data. To assess fidelity, we can use the cumulative distribution function (CDF), following the approach proposed by [160]. 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 [10, 19, 34], 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. [48] 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 [99, 121, 160, 145].

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 [109]. 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 [104]. Class imbalance, in particular, has been shown to result in unfair decisions for minority classes, as highlighted in [25].

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 [146], where augmented datasets were used to enhance model performance and fairness.

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

The effectiveness of each ML model will be 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. 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 Iterative 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 [16, 122, 140] and provide effective benchmarks for assessing the performance of our algorithm.

## 6.1 Algorithm UMAP-PSDG. Fidelity

In the previous section, the proposed methodology was applied to generate synthetic data using the data sets introduced in Section 3. Since the true values for the missing data in the incomplete data sets are not available, direct fidelity validation of the generated synthetic data cannot be performed. Now, as an alternative, a validation process is carried out using the PI-CAI data set for fidelity. This involves partitioning PI-CAI into reference and incomplete data sets, allowing the generated synthetic values to be compared against the original values. This comparison facilitates a thorough evaluation of the performance and accuracy of the proposed methodology.

To perform the validation, the complete and known PI-CAI database, with 1500 samples, was divided into five random groups, each containing 300 samples. One group was used as the reference data set (PI-CAI 1), while the other four groups served as incomplete data (PI-CAI 2). This process was repeated until all the data groups were utilized as reference data.

Using PI-CAI 1 as the reference data and PI-CAI 2 as the incomplete data, we applied our previously described methodology. By comparing the generated samples with the real data available in PI-CAI 2, we can evaluate the accuracy of the algorithm in imputing missing values and generating synthetic samples.

For each iteration, the methodology was applied, generating validated data. The validated data contained synthetic samples that did not correspond to any real patient and also correctly imputed data, by imputing the correct missing value of the incomplete data. After each iteration, the number of correctly imputed and synthetic generated data was recorded and averaged across the five iterations. This averaging process provides a more stable estimation of the performance of our algorithm.

In Figure 10, the amount of correctly imputed data is represented for each reliability threshold. As a reference, the number of samples originally corresponding to each age in the reference data set (PI-CAI 2) are represented in gray. Then, the number of correctly imputed samples per age are represented for the different reliability thresholds. For a more numerical interpretation, the percentage of correctly imputed data for certain age ranges is shown in Table 4. As expected, the number of correctly imputed samples increases as the reliability threshold decreases.

All the validated samples that do not correspond to real data are synthetic samples. In Figure 11, the number of generated samples per age depending on the reliability threshold can be observed. It is observed that the number of samples is augmented significantly. Also, as expected, ages with lower representation result in a lower number of generated samples.

For lower reliability values, a higher number of generated data is obtained. From the generated data, a percentage between 3.49% ( $r_{sc} = 1$ ) and 2.79% ( $r_{sc} = 0.5$ ) corresponds to correctly imputed data, the rest of the generated data are synthetic samples.

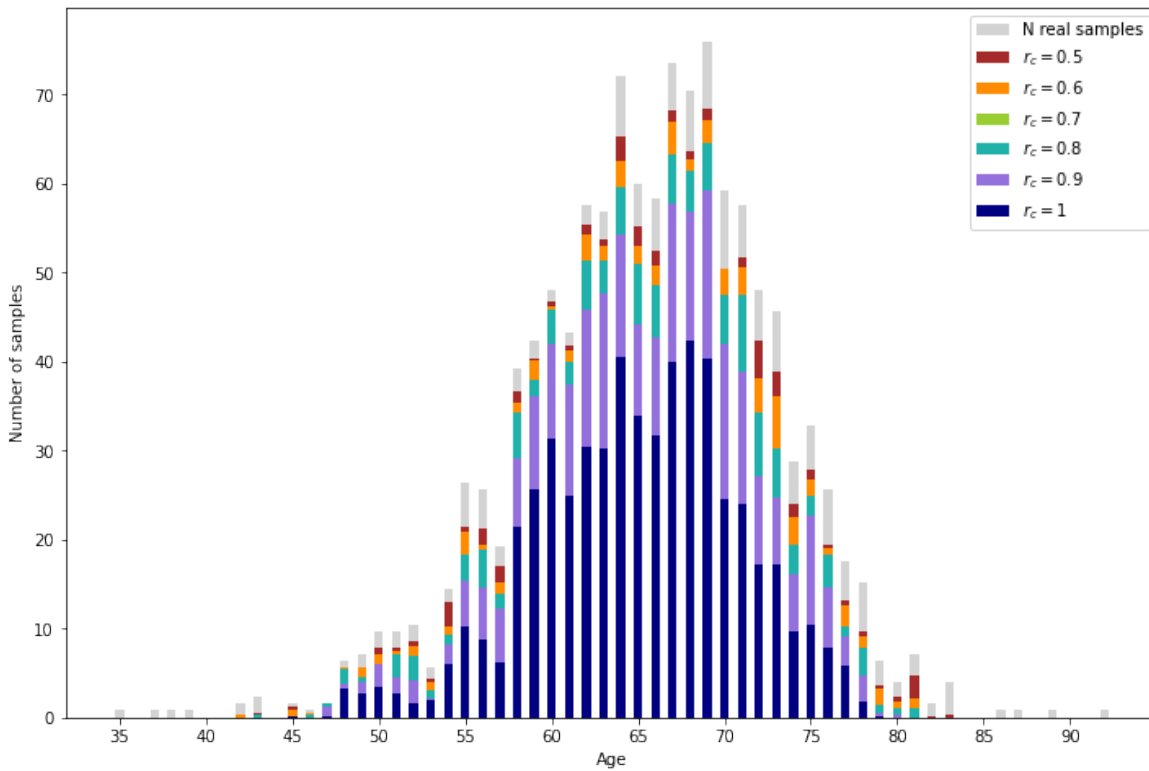


Figure 10: Real samples from the incomplete database (light gray) and correctly imputed samples for different reliability values.

Aside from the usability of the synthetic samples, the proposed methodology presents a higher imputation performance than the typical methods such as mean or  $k$ -NN imputation. As can be observed in Figure 12, the mean imputation, which assigns the mean value of the reference data for the missing variable to all the missing samples, only obtains a low percentage of correct imputation for the resulting mean ages. On the other hand,  $k$ -NN imputation calculates the mean values of the missing variable based on the nearest neighbors, identified by computing the Euclidean distance between the feature vectors of the samples. It can be observed that there are more ages with correctly imputed samples than with mean imputation, but the amount of correctly imputed samples is very low. Finally, by applying the proposed methodology with the most restrictive reliability score  $r_{sc} = 1$ , the number of correctly imputed samples increases significantly. A more detailed perspective can be obtained in Table 4 by comparing the exact percentage of correctly imputed samples.

Given that traditional methods lack the capability to leverage complete data as reference information, a direct comparison with the proposed methodology is not entirely appropriate. However, for a general sense of performance, it is worth noting that even the most restrictive reliability score in our method results in a validated sample percentage above 25% while the traditional methods only reach 0.45% and 2.02%.

Also, a possible criterion to determine the reliability score is to consider the percentage of correctly imputed samples from the generated data. It can be seen that the percentage does not change between  $r_c = 0.7$  and  $r_c = 0.8$ . However, the largest drop in the percentage occurs between  $r_c = 0.9$  and  $r_c = 1$ .

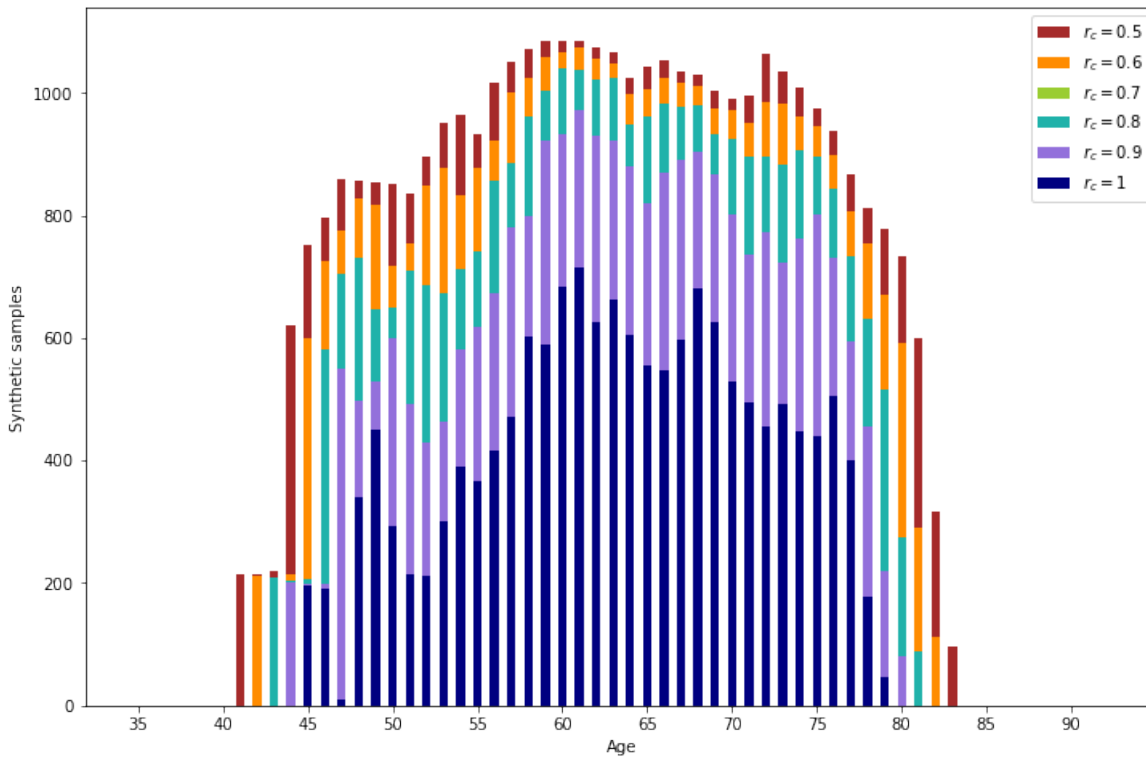


Figure 11: Number of synthetic samples per age.

Hence, for the validation against the PI-CAI data set, the percentage of correctly imputed values underscores the effectiveness of the proposed methodology, particularly when compared to traditional imputation techniques such as mean imputation or  $k$ -nearest neighbors ( $k$ -NN) imputation.

## 6.2 Algorithm UMAP-PSDG. Privacy and workflow strategies

The proposed methodology is designed not only to augment or complete data sets, but also to ensure data privacy when the reference and incomplete data sets are located in different data centers. While the previous section detailed the methodology for general reference and incomplete data sets, this section focuses on the information workflow between separate data centers, emphasizing the protection of sensitive information.

To illustrate this, we revisit the scenario involving hospital 1 and hospital 2. In this case, hospital 2 ( $H_2$ ) possesses a large but incomplete data set and seeks to leverage the complete data owned by hospital 1 ( $H_1$ ) to fill in the missing information. Our methodology offers a solution to this challenge by enabling data exchange without compromising privacy, ensuring that sensitive data remain secure throughout the process.

Three scenarios are considered depending on the unknown information in the incomplete data set in  $H_2$ :

- Case 0: The output is unknown, but the features are all known. In the case of missing outputs or classes, the validation process involves cluster validation, which requires the centroid cluster coordinates for the reference dataset and a threshold radius from  $H_1$ .

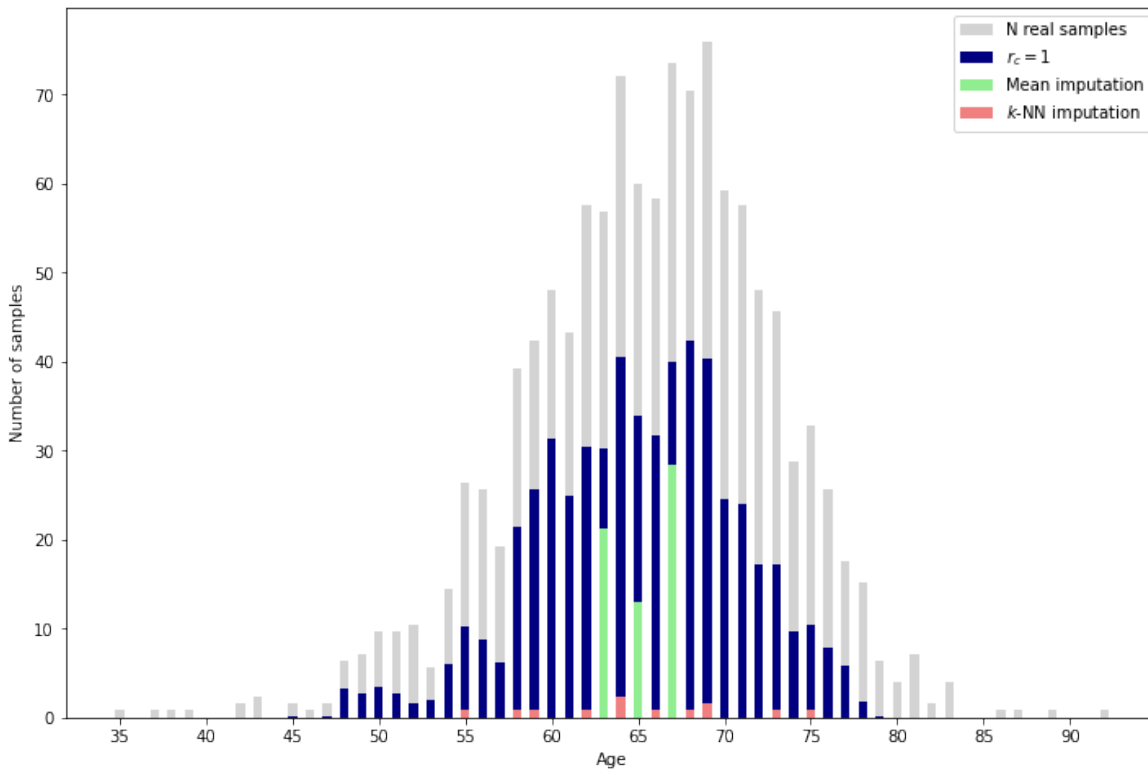


Figure 12: Imputation performance of mean and  $k$ -NN ( $k = 10$ ) vs. proposed methodology.

- Case 1: The output is known, but there is one missing feature. If the interest lies in assigning generated values for a missing feature, the validation process involves a reliability score assignment. This requires the mean value of the missing feature for the  $k$ -nearest neighbors in the coordinates of the reference dataset, for each instance  $i$ .
- Case 2: Both, the output and a feature are unknown. For scenarios involving the generation of both the unknown feature and the assignment of sample clusters, the validation process includes both steps.

Regardless of the scenario, all workflows require the UMAP function trained with the reference data in  $H_1$ .

Our primary interest lies in Case 1, which results in a synthetic feature for each entry. The other cases are presented for a better understanding of the system, but the workflow proposal focuses on Case 1.

Depending on the interests and needs of the health institutions and their data centers, there exist different workflow options for sharing information, shown in Figure 13. The UMAP function,  $f_{umap,r}$ , trained with the data in  $H_1$ , must be shared with hospital 2. This step does not pose any privacy risks because the reference data cannot be obtained back from the trained function.

The second information exchange occurs when hospital 2 transmits the UMAP coordinates of the generated dataset to the reference institution, hospital 1. Notably, in this exchange,  $H_1$  does not obtain sensitive high-dimensional data due to the inherent limitations of UMAP coordinates. Since both the UMAP transformation, and its inverse, are stochastic processes,

Table 4: Percentage of correctly imputed samples with the proposed methodology, mean and  $k$ -NN imputation.

Age Range	35–44	45–54	55–64	65–74	75–84	85–92	Mean %
$r_{sc} = 0.5$	16.67	83.93	91.73	88.09	66.26	0.00	57.78
$r_{sc} = 0.6$	11.11	78.57	89.03	85.11	59.09	0.00	53.82
$r_{sc} = 0.7$	5.56	70.24	84.99	80.19	52.97	0.00	48.99
$r_{sc} = 0.8$	5.56	70.24	84.99	80.19	52.97	0.00	48.99
$r_{sc} = 0.9$	0.00	54.46	76.25	71.30	41.61	0.00	40.60
$r_{sc} = 1.0$	0.00	33.04	53.39	48.68	22.73	0.00	26.31
Mean	0	0	4.93	7.17	0	0	2.02
$k$ -NN	0	0	1.30	0.69	0.70	0	0.45

noise inevitably accompanies every transformation. Consequently,  $H_1$  can only approximate the generated data from  $H_2$  by inversely transforming the shared coordinates. The information exchange in this step is crucial for computing the mean value of the missing feature for the  $k$  neighbors in the reference coordinates for each generated sample.

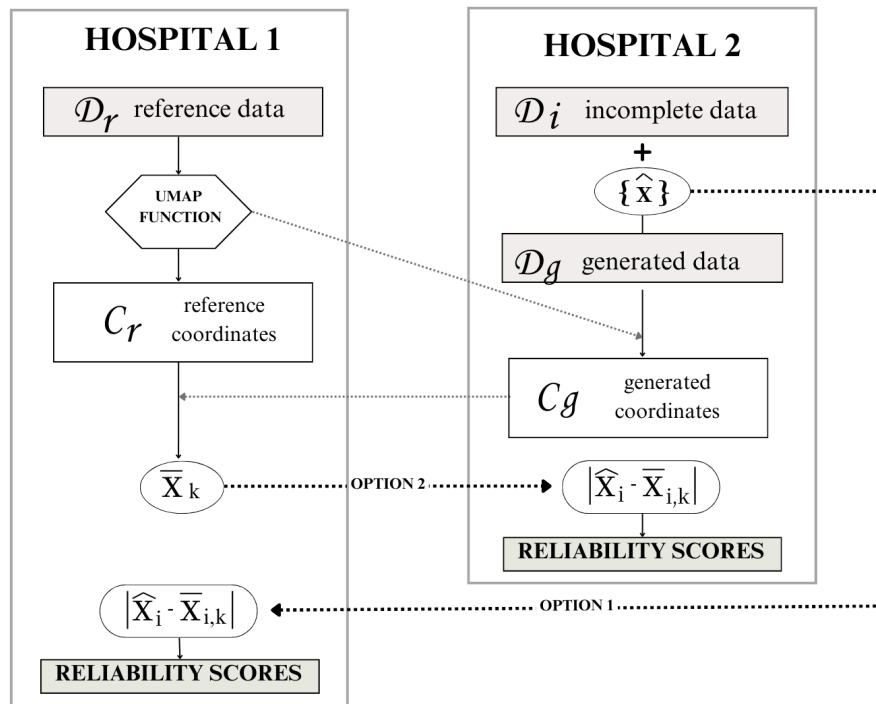


Figure 13: Workflow options according to data privacy.

Finally, to perform the reliability score assignment, there are two options:

- **Option 1:** Hospital 2 shares the generated value for each sample with hospital 1. Hospital 1 calculates the feature disparity and assigns a reliability score to each generated sample. Hospital 1 then sends the reliability scores back to hospital 2 for them to keep the desired samples.

- **Option 2** Hospital 1 shares the mean neighbor value of the missing feature for each sample with hospital 2. Hospital 2 calculates the feature disparity to provide the reliability scores.

## 6.3 Algorithm Iterative UMAP-FSDG. Fidelity and utility

After generating the synthetic datasets, a quality assessment is conducted to evaluate the effectiveness of the Iterative UMAP-based approach. To establish a performance benchmark, we compare our results to 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 breast 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 represents 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.

### 6.3.1 Data fidelity

The fidelity assessment of the synthetic data generation through CDF is done both visually 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 5, provides a robust metric for evaluating distributional similarity for each attribute.

The quantitative results of the fidelity assessment, as measured by the Kolmogorov-Smirnov (K-S) statistic, are presented in Table 5, 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 achieves the lowest K-S statistic for most attributes in the breast cancer dataset, emphasizing its ability to closely replicate the real data distributions in this scenario.

### 6.3.2 Data utility

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.

Table 5: 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	<b>0.1554</b>	0.1673
	psa	0.4103	0.4695	0.3583	<b>0.1980</b>
	prostate_volume	0.3639	0.3557	0.1690	<b>0.1327</b>
Breast Cancer	age	<b>0.3262</b>	0.3990	0.3289	0.5524
	tumor-size	0.1793	<b>0.1344</b>	0.2042	0.1410
	inv-nodes	<b>0.6010</b>	0.6942	0.6254	0.9477
	node-caps	<b>0.4254</b>	0.7501	0.6662	0.8906
	deg-malig	<b>0.3756</b>	0.4090	0.3876	0.5921
Cardiov. Disease	irradiat	<b>0.3931</b>	0.7118	0.6330	0.9448
	chestpain	<b>0.3505</b>	0.4443	0.3834	0.4865
	restingBP	0.1351	<b>0.0382</b>	0.1419	0.1478
	fastingbloodsugar	<b>0.5749</b>	0.6689	0.6730	0.8934
	maxheartrate	0.0568	<b>0.0404</b>	0.1919	0.3230

**Results for Random Forest classifier.** For the Random Forest Classifier, the performance metrics are presented in Figure 14. For the prostate cancer data (see Figure 14a), the UMAP-generated and TVAE synthetic and augmented datasets show superior performance compared to training on real data alone or training with CTGAN and CopulaGAN synthetic and augmented datasets [96].

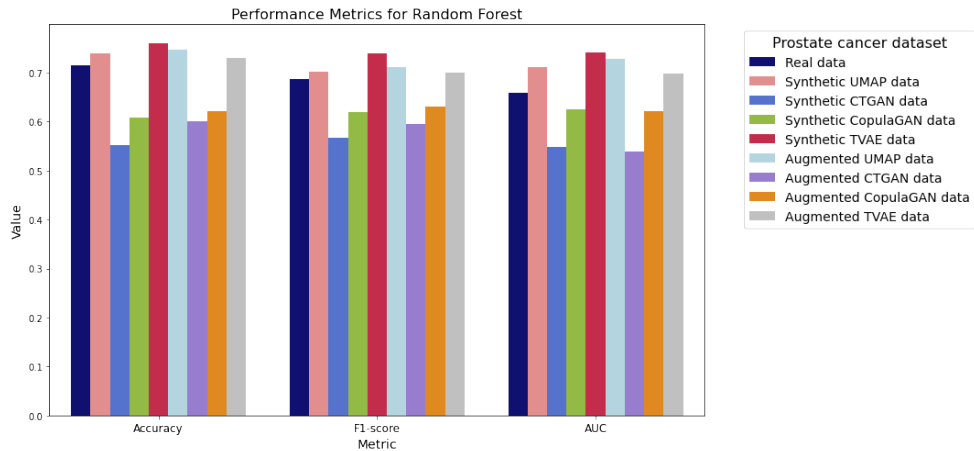
For the breast cancer data (see Figure 14b), the performance enhancement provided by TVAE is less pronounced. However, UMAP synthetic and augmented datasets exhibit better results than real data across all metrics, except for AUC. In this specific case, only the augmented UMAP and TVAE datasets match the AUC performance achieved by real data [96].

In the case of the cardiovascular data (see Figure 14c), none of the synthetic or augmented datasets surpass the performance of the real dataset. Additionally, CTGAN-generated dataset demonstrates particularly poor performance metrics for this scenario [96].

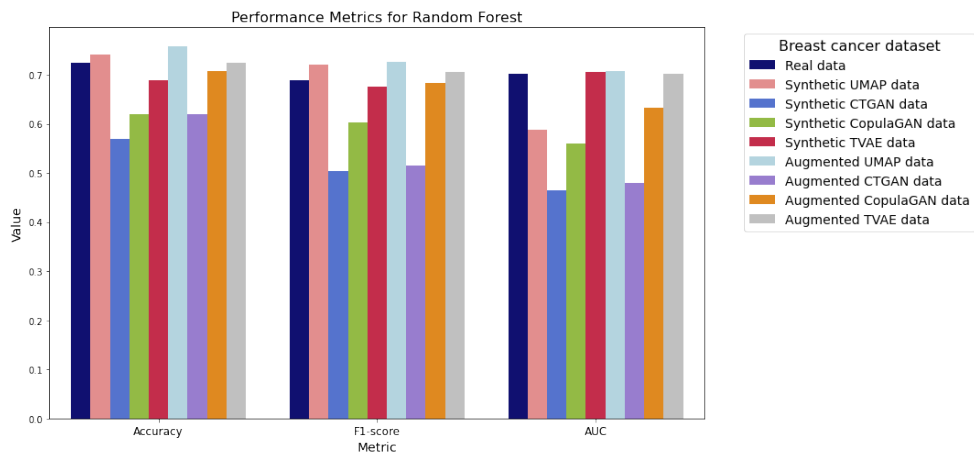
**Results for logistic regression models.** The performance metrics for Logistic Regression models, as displayed in Figure 15, reveal distinct patterns across the datasets. In the prostate cancer case (see Figure 15a), while UMAP and TVAE-generated datasets outperform CTGAN and CopulaGAN, they do not consistently exceed the performance of models trained on real data.

For breast cancer data (see Figure 15b), CopulaGAN-generated datasets demonstrate improved performance compared to previous observations, particularly in the F1-score metric. Additionally, both CopulaGAN and UMAP-generated datasets slightly outperform the real dataset in accuracy.

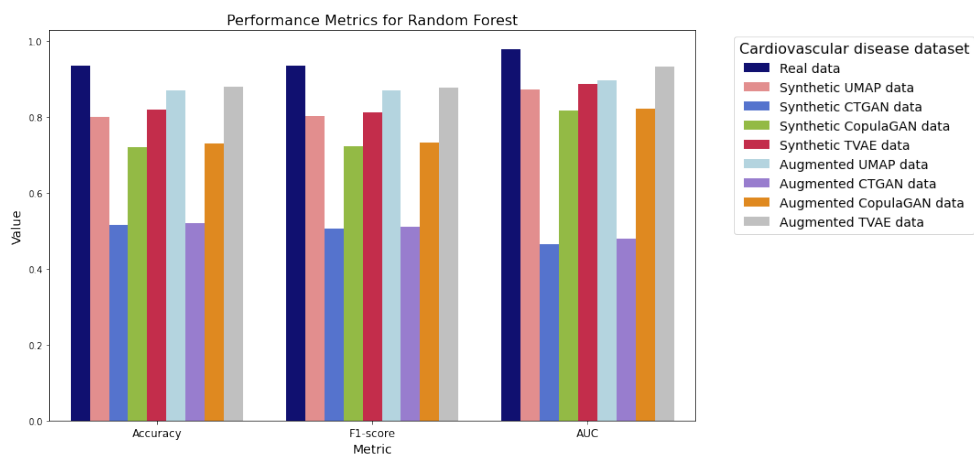
In the cardiovascular diseases scenario (see Figure 15c), similar to the Random Forest results, none of the synthetic datasets outperform the model trained with real data.



(a) Prostate cancer data



(b) Breast cancer data

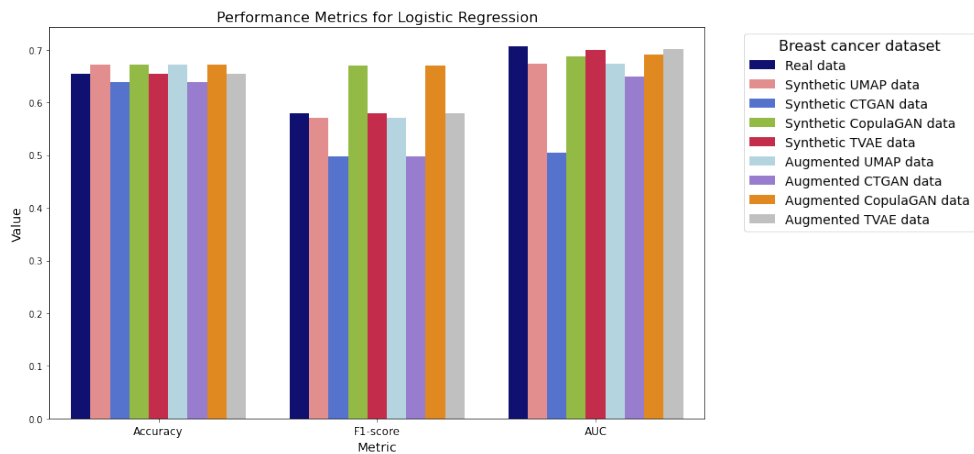


(c) Cardiovascular diseases data

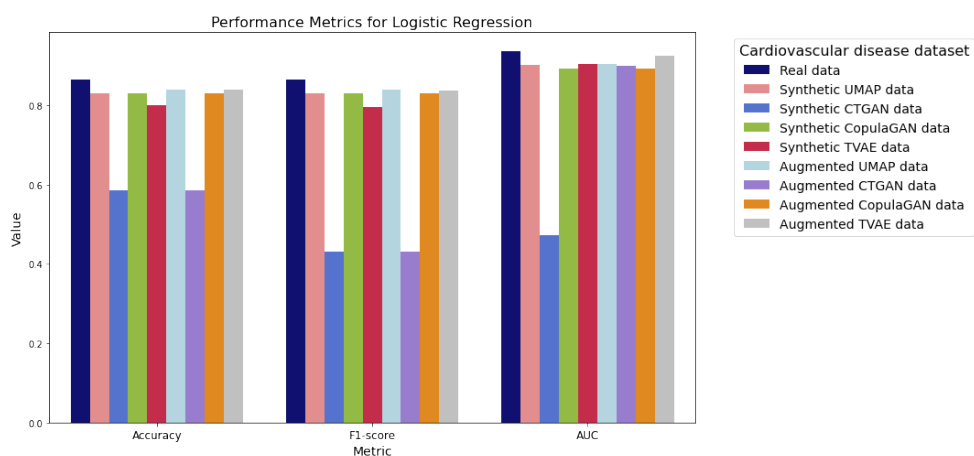
Figure 14: Utility metrics (Accuracy, F1-score and AUC) for Random Forest model trained with real, synthetic and augmented data



(a) Prostate cancer data



(b) Breast cancer data



(c) Cardiovascular diseases data

Figure 15: Utility metrics (Accuracy, F1-score and AUC) for Logistic Regression model trained with real, synthetic and augmented data

**Results for Support Vector Machines** Support Vector Machine (SVM), presented in Figure 16, results provide further insights. For prostate cancer data (see Figure 16a), 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 (see Figure 16b), 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 shows a marked enhancement in AUC performance.

Contrasting with the previous models, the cardiovascular diseases scenario (see Figure 16c) showcases superior performance for synthetic and augmented datasets compared to real data. Among these, the UMAP-generated data stands out, delivering excellent results and outperforming all other datasets across the evaluated metrics.

### 6.3.3 Discussion of results

With this algorithm, 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 Iterative 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 to 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.

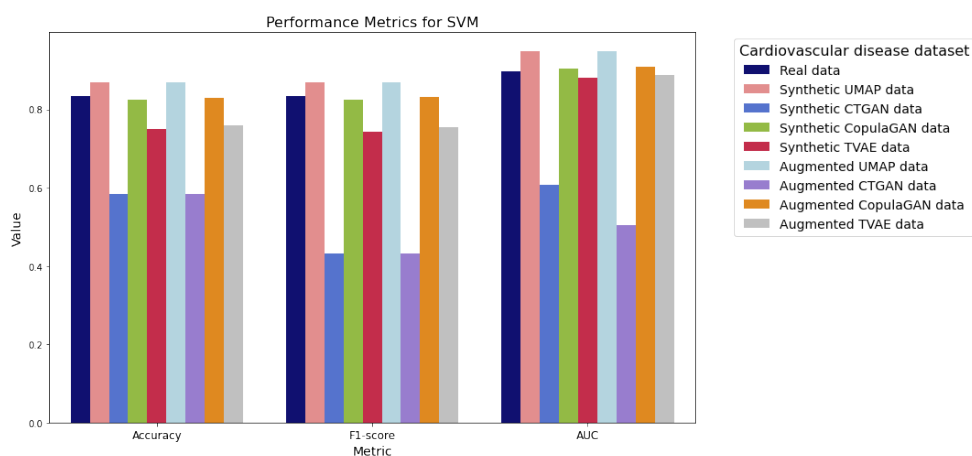
As highlighted in the results analysis, augmented datasets often yield better utility outcomes compared to real or synthetic datasets alone. Data augmentation serves as a powerful tool



(a) Prostate cancer data



(b) Breast cancer data



(c) Cardiovascular diseases data

Figure 16: Utility metrics (Accuracy, F1-score and AUC) for Support Vector Machine model trained with real, synthetic and augmented data

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.

## 7 Conclusions and Future Work

Tasks associated to the deliverable *D3.1. Structured synthetic health data generation* include the development of generative tools for structured (tabular) synthetic data generation that lead to good results from the statistical evaluation perspective. Synthetic data generation will be performed on the cloud-based FLUTE platform, as part of the services provided by the platform to researchers and innovators. Hence, the aim of our research is primarily to develop a

methodology for generating, partially or fully, synthetic data.

Initially, checking the reality of our clinical partners, developments started on partially synthetic data generation: from an incomplete database by leveraging a complete database as reference data, while addressing privacy concerns. The proposed methodology successfully imputed missing data and generated synthetic samples, surpassing the number of incomplete data entries. In this form, the methodology provides a secure framework for data augmentation by utilizing data from different centers without the need to transfer sensitive information. Additionally, it yields superior results for data imputation tasks. The developed methodology has practical applications in generating partially synthetic data that do not correspond to any specific patient.

Next, as discussed throughout the deliverable, the partially synthetic data generation method is adapted to produce fully synthetic data sets. Consequently, the proposed method can be applied to any data generation scenario. This has significant implications for data-driven research and decision making in healthcare and other fields. Moreover, this is the first time that visualization techniques and dimension reduction methods are employed in this domain.

The proposed novel methodology represents a valuable tool for generating synthetic data, providing a balance between data utility and privacy preservation. Further research can explore its application in other domains and investigate additional methods for enhancing its performance and scalability. In particular, testing alternative dimensionality reduction methods across diverse data sets with varying characteristics and higher dimensionalities would provide valuable insights into its adaptability and effectiveness. Additionally, by refining the methodology for fully synthetic data generation, its outcomes can be more directly compared to established methods, facilitating a comprehensive evaluation of its potential across fields.

In conclusion, this method represents a robust solution for generating secure, high-quality synthetic healthcare data, effectively addressing data scarcity challenges.

Beyond the algorithms, the task is also focused on the quality of the generated synthetic data. As it has been demonstrated, the generated synthetic data has been rigorously evaluated for fidelity and utility measures. Results show that the UMAP-based algorithm outperforms GAN and VAE-based generation methods across different scenarios. In fidelity assessments, it achieves 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.

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, a future research work can involve testing different dimensionality reduction tools. Although UMAP has shown an excellent performance, specially for health-care 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 will 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.

As a secondary result of this task, we have provided a formal representation for the synthetic data generation problem, specifically addressing cases of partially synthetic data generation. This objective has been successfully achieved, as the work articulates a clear and formal definition of the problem, thereby contributing to the ongoing discourse in the field of data synthesis.

## References

- [1] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*, pages 510–526. Springer, 2019.
- [2] Nasrullah Abbasi, FNU Nizamullah, Shah Zeb, and MD Fardous. Generative ai in healthcare: Revolutionizing disease diagnosis, expanding treatment options, and enhancing patient care. *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, 3(3):127–138, 2024.
- [3] Ahmed Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 290–306. PMLR, 17–23 Jul 2022.

- [4] Goncalo Almeida and Fernando Bacao. Umap-smotenc: A simple, efficient, and consistent alternative for privacy-aware synthetic data generation. *Knowledge-Based Systems*, page 112174, 2024.
- [5] Ehsan Amid and Manfred K Warmuth. Trimap: Large-scale dimensionality reduction using triplets. *arXiv preprint arXiv:1910.00204*, 2019.
- [6] Patricia A. Apellániz, Ana Jim’enez, Borja Arroyo Galende, Juan Parras, and Santiago Zazo. Synthetic tabular data validation: A divergence-based approach. *IEEE Access*, 12:103895–103907, 2024.
- [7] Patricia A Apellániz, Juan Parras, and Santiago Zazo. An improved tabular data generator with vae-gmm integration. *arXiv preprint arXiv:2404.08434*, 2024.
- [8] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.
- [9] Marc Ayoub, Ahmad A Ballout, Rosana A Zayek, and Noel F Ayoub. Mind+ machine: Chatgpt as a basic clinical decisions support tool. *Cureus*, 15(8), 2023.
- [10] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association*, 26(3):228–241, 2019.
- [11] André Bauer, , Simon Trapp, , Michael Stenger, , Robert Leppich, , Samuel Kounev, , Mark Leznik, , Kyle Chard, , and Ian Foster. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint*, page 103, 2024. <https://doi.org/10.48550/arXiv.2401.02524>.
- [12] Brett K Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, and Casey S Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019.
- [13] Ana Beduschi. Synthetic data protection: Towards a paradigm change in data regulation? *Big Data & Society*, 11(1):20539517241231277, 2024.
- [14] Karan Bhanot, Miao Qi, John S Erickson, Isabelle Guyon, and Kristin P Bennett. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165, 2021.
- [15] Harsh Bhatt, Nilesh Kumar Jadav, Aparna Kumari, Rajesh Gupta, Sudeep Tanwar, Zdzislaw Polkowski, Amr Tolba, and Azza S Hassanein. Artificial neural network-driven federated learning for heart stroke prediction in healthcare 4.0 underlying 5g. *Concurrency and Computation: Practice and Experience*, 36(3):e7911, 2024.
- [16] Dominik Bietsch, Robert Stahlbock, and Stefan Voß. Synthetic data as a proxy for real-world electronic health records in the patient length of stay prediction. *Sustainability*, 15(18):13690, 2023.
- [17] Rohit Bokade, Alfred Navato, Ruilin Ouyang, Xiaoning Jin, Chun-An Chou, Sarah Ostadabbas, and Amy V Mueller. A cross-disciplinary comparison of multimodal data

- fusion approaches and applications: Accelerating learning through trans-disciplinary information sharing. *Expert Systems with Applications*, 165:113885, 2021.
- [18] Alexander Boudewijn, Andrea Filippo Ferraris, Daniele Panfilo, Vanessa Cocca, Sabrina Zinutti, Karel De Schepper, and Carlo Rossi Chauvenet. Privacy measurement in tabular synthetic data: State of the art and future research directions, 2023.
- [19] Stavroula Bourou, Andreas El Saer, Terpsichori-Helen Velivassaki, Artemis Voulkidis, and Theodore Zahariadis. A review of tabular data synthesis using gans on an ids dataset. *Information*, 12(09):375, 2021.
- [20] Tim Bradshaw and Leila Abboud. Financial Times. Four-week-old AI start-up raises record €105mn in european push. <https://www.ft.com/content/cf939ea4-d96c-4908-896a-48a74381f251>, 2023. Access Jun-20-2023.
- [21] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, and Chidchanok Lursinsap. Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Advances in knowledge discovery and data mining: 13th Pacific-Asia conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 proceedings 13*, pages 475–482. Springer, 2009.
- [22] Qiong Cai, Hao Wang, Zhenmin Li, and Xiao Liu. A survey on multimodal data-driven smart healthcare systems: approaches and applications. *IEEE Access*, 7:133583–133599, 2019.
- [23] Ramiro Camino, Christian Hammerschmidt, and Radu State. Generating multi-categorical samples with generative adversarial networks, 2018.
- [24] Ramiro Camino, Christian Hammerschmidt, and Radu State. Generating multi-categorical samples with generative adversarial networks. *arXiv preprint arXiv:1807.01202*, 2018.
- [25] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.
- [26] Pavitra Chauhan, Lars Ailo Bongo, and Edvard Pedersen. Ethical challenges of using synthetic data. In *Proceedings of the AAAI Symposium Series*, volume 1(1), pages 133–134, 2023.
- [27] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, jun 2002.
- [28] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1691–1703. PMLR, 13–18 Jul 2020.
- [29] Yan Chen and Pouyan Esmaeilzadeh. Generative ai in medical practice: in-depth exploration of privacy and security challenges. *Journal of Medical Internet Research*, 26:e53008, 2024.

- [30] Kieran Chin-Cheong, Thomas Sutter, and Julia E Vogt. Generation of heterogeneous synthetic electronic health records using gans. In *workshop on machine learning for health (ML4H) at the 33rd conference on neural information processing systems (NeurIPS 2019)*. ETH Zurich, Institute for Machine Learning, 2019.
- [31] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jiemeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens, editors, *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR, 18–19 Aug 2017.
- [32] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jiemeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR, 2017.
- [33] João Coutinho-Almeida, Pedro Pereira Rodrigues, and Ricardo João Cruz-Correia. Gans for tabular healthcare data generation: A review on utility and privacy. In *Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings*, page 282–291, Berlin, Heidelberg, 2021. Springer-Verlag.
- [34] Fida K Dankar, Mahmoud K Ibrahim, and Leila Ismail. A multi-dimensional evaluation of synthetic data generators. *IEEE Access*, 10:11147–11158, 2022.
- [35] Salman UH. Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Çukur. Image synthesis in multi-contrast mri with conditional generative adversarial networks. *IEEE Transactions on Medical Imaging*, 38(10):2375–2388, 2019.
- [36] Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.
- [37] Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. *arXiv preprint arXiv:1802.04208*, 2018.
- [38] Jörg Drechsler and Anna-Carolina Haensch. 30 years of synthetic data. *Statistical Science*, 39(2):221–242, 2024.
- [39] Junwei Duan, Jiaqi Xiong, Yinghui Li, and Weiping Ding. Deep learning based multi-modal biomedical data fusion: An overview and comparative review. *Information Fusion*, page 102536, 2024.
- [40] Tri Dung Duong, Qian Li, and Guandong Xu. Ceflow: A robust and efficient counterfactual explanation framework for tabular data using normalizing flows. In Hisashi Kashima, Tsuyoshi Ide, and Wen-Chih Peng, editors, *Advances in Knowledge Discovery and Data Mining*, pages 133–144, Cham, 2023. Springer Nature Switzerland.
- [41] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.
- [42] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Ap-*

- plications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, pages 486–503. Springer, 2006.
- [43] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.
- [44] Peter Eigenschink, Thomas Reutterer, Stefan Vamosi, Ralf Vamosi, Chang Sun, and Klaudius Kalcher. Deep generative models for synthetic data: A survey. *IEEE Access*, 11:47304–47320, 2023.
- [45] K.E. Emam, L. Mosquera, and R. Hoptroff. *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. O’Reilly Media, 2020.
- [46] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 1054–1067, 2014.
- [47] Erica Espinosa and Alvaro Figueira. On the quality of synthetic generated tabular data. *Mathematics*, 11(15):3278, 2023.
- [48] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.
- [49] Giovanni Fasano and Alberto Franceschini. A multidimensional version of the kolmogorov–smirnov test. *Monthly Notices of the Royal Astronomical Society*, 225(1):155–170, 1987.
- [50] Stefan Feuerriegel, Jochen Hartmann, Christian Janiesch, and Patrick Zschech. Generative ai. *Business & Information Systems Engineering*, 66(1):111–126, 2024.
- [51] Alvaro Figueira and Bruno Vaz. Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, 10(15), 2022.
- [52] Joao Fonseca and Fernando Bacao. Tabular and latent space synthetic data generation: a literature review. *Journal of Big Data*, 10(1):115, 2023.
- [53] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.
- [54] Nir Friedman and Daphne Koller. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. *Machine learning*, 50:95–125, 2003.
- [55] Oscar Giles, Kasra Hosseini, Grigorios Mingas, Oliver Strickson, Louise Bowler, Camila Rangel Smith, Harrison Wilde, Jen Ning Lim, Bilal Mateen, Kasun Amarasinghe, et al. Faking feature importance: A cautionary tale on the use of differentially-private synthetic data. *arXiv preprint arXiv:2203.01363*, 2022.

- [56] Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. A unified framework for quantifying privacy risk in synthetic data, 2022.
- [57] Mauro Giuffrè and Dennis L Shung. Harnessing the power of synthetic data in health-care: innovation, application, and privacy. *NPJ digital medicine*, 6(1):186, 2023.
- [58] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. Generation and evaluation of synthetic patient data. *BMC medical research methodology*, 20:1–40, 2020.
- [59] Aldren Gonzales, Guruprabha Guruswamy, and Scott R Smith. Synthetic data in health care: a narrative review. *PLOS Digital Health*, 2(1):e0000082, 2023.
- [60] Luis Gonzalez-Abril, Cecilio Angulo, Juan-Antonio Ortega, and José-Luis Lopez-Guerra. Generative adversarial networks for anonymized healthcare of lung cancer patients. *Electronics*, 10(18), 2021.
- [61] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [62] Wenzhong Guo, Jianwen Wang, and Shiping Wang. Deep multimodal representation learning: A survey. *Ieee Access*, 7:63373–63394, 2019.
- [63] Aman Gupta, Deepak Bhatt, and Anubha Pandey. Transitioning from real to synthetic data: Quantifying the bias in model, 2021.
- [64] Hansle Gwon, Imjin Ahn, Yunha Kim, Hee Jun Kang, Hyeram Seo, Heejung Choi, Ha Na Cho, Minkyoung Kim, JiYe Han, Gaeun Kee, et al. Ldp-gan: Generative adversarial networks with local differential privacy for patient medical records synthesis. *Computers in Biology and Medicine*, 168:107738, 2024.
- [65] Muhammad Salman Haleem, Audrey Ekuban, Alessio Antonini, Silvio Pagliara, Leandro Pecchia, and Carlo Allocca. Deep-learning-driven techniques for real-time multimodal health and physical data synthesis. *Electronics*, 12(9):1989, 2023.
- [66] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.
- [67] Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity score: A new metric to evaluate the uncommonness of synthesized images. *arXiv preprint arXiv:2206.08549*, 2022.
- [68] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008.
- [69] Huan He, Shifan Zhao, Yuanzhe Xi, and Joyce C Ho. Meddiff: Generating electronic health records using accelerated denoising diffusion model. *arXiv preprint arXiv:2302.04355*, 2023.

- [70] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic tabular data evaluation in the health domain covering resemblance, utility, and privacy dimensions. *Methods of information in medicine*, 62(S 01):e19–e38, 2023.
- [71] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing*, 493:28–45, 2022.
- [72] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. Standardised metrics and methods for synthetic tabular data evaluation. *Authorea Preprints*, 2023.
- [73] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
- [74] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [75] Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, 28(1):65–98, 2019.
- [76] Kai Hu, Sheng Gong, Qi Zhang, Chaowen Seng, Min Xia, and Shanshan Jiang. An overview of implementing security and privacy in federated learning. *Artificial Intelligence Review*, 57(8):204, 2024.
- [77] Nathan C Hurley, Adrian D Haimovich, R Andrew Taylor, and Bobak J Mortazavi. Visualization of emergency department clinical data for interpretable patient phenotyping. *Smart Health*, 25:100285, 2022.
- [78] Tasin Islam, Alina Miron, Monomita Nandy, Jyoti Choudrie, Xiaohui Liu, and Yongmin Li. Transforming digital marketing with generative ai. *Computers*, 13(7):168, 2024.
- [79] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. Synthetic data – what, why and how?, 2022.
- [80] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.
- [81] Bahador Khaleghi, Alaa Khamis, Fakhreddine O Karray, and Saiedeh N Razavi. Multisensor data fusion: A review of the state-of-the-art. *Information fusion*, 14(1):28–44, 2013.
- [82] Shivansh Khanna and Shraddha Srivastava. Ai governance in healthcare: Explainability standards, safety protocols, and human-ai interactions dynamics in contemporary medical ai systems. *Empirical Quests for Management Essences*, 1(1):130–143, 2021.
- [83] Dong-Keon Kim, DongHeum Ryu, Yongbin Lee, and Dong-Hoon Choi. Generative models for tabular data: A review. *Journal of Mechanical Science and Technology*, pages 1–17, 2024.

- [84] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [85] Ivan Kobyzev, Simon Prince, and Marcus A Brubaker. Normalizing flows: Introduction and ideas. *stat*, 1050:25, 2019.
- [86] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*, pages 17564–17579. PMLR, 2023.
- [87] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.
- [88] Abdullah Lakhan, Hassen Hamouda, Karrar Hameed Abdulkareem, Saleh Alyahya, and Mazin Abed Mohammed. Digital healthcare framework for patients with disabilities based on deep federated learning schemes. *Computers in Biology and Medicine*, 169:107845, 2024.
- [89] Dongha Lee, Hwanjo Yu, Xiaoqian Jiang, Deevakar Rogith, Meghana Gudala, Mubeen Tejani, Qiuchen Zhang, and Li Xiong. Generating sequential electronic health records using dual adversarial autoencoder. *Journal of the American Medical Informatics Association*, 27(9):1411–1419, 2020.
- [90] Jaewoo Lee, Minjung Kim, Yonghyun Jeong, and Youngmin Ro. Differentially private normalizing flows for synthetic tabular data generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7345–7353, 2022.
- [91] Qing Li, Guanyuan Yu, Jun Wang, and Yuehao Liu. A deep multimodal generative and fusion framework for class-imbalanced multimodal data. *Multimedia Tools and Applications*, 79:25023–25050, 2020.
- [92] RuiBo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinqing Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, et al. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*, 2024.
- [93] Anton Lysenko, Irina Deeva, and Egor Shikov. MvaeSynth: a unified framework for multimodal data generation, modality restoration, and controlled generation. *Procedia Computer Science*, 193:422–431, 2021.
- [94] Lingjuan Lyu, Han Yu, Xingjun Ma, Chen Chen, Lichao Sun, Jun Zhao, Qiang Yang, and S Yu Philip. Privacy and robustness in federated learning: Attacks and defenses. *IEEE transactions on neural networks and learning systems*, 2022.
- [95] Lingjuan Lyu, Han Yu, and Qiang Yang. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*, 2020.
- [96] Carla Lázaro and Cecilio Angulo. Iterative application of umap-based algorithms for fully synthetic healthcare tabular data generation. *Algorithms*, 17(12), 2024.
- [97] Carla Lázaro and Cecilio Angulo. Using umap for partially synthetic healthcare tabular data generation and validation. *Sensors*, 24(23), 2024.

- [98] David JC MacKay et al. Introduction to gaussian processes. *NATO ASI series F computer and systems sciences*, 168:133–166, 1998.
- [99] Manjunath Mahendra, Chaithra Umesh, Saptarshi Bej, Kristian Schultz, and Olaf Wolkenhauer. Convex space learning for tabular synthetic data generation. *arXiv preprint arXiv:2407.09789*, 2024.
- [100] Stan Matwin, Jordi Nin, Morvarid Sehatkar, and Tomasz Szapiro. A review of attribute disclosure control. *Advanced research in data privacy*, pages 41–61, 2015.
- [101] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [102] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [103] Ofer Mendeleevitch and Michael D. Lesh. Fidelity and privacy of synthetic medical data. *CoRR*, abs/2101.08658, 2021.
- [104] Nicolo Micheletti, Raffaele Marchesi, Nicholas I-Hsien Kuo, Sebastiano Barbieri, Giuseppe Jurman, and Venet Osmani. Generative ai mitigates representation bias and improves model fairness through synthetic health data. *medRxiv*, pages 2023–09, 2023.
- [105] Muzafar Mehraj Misgar and MPS Bhatia. Detection of depression from iomt time series data using umap features. In *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 623–628. IEEE, 2022.
- [106] Ghulam Muhammad, Fatima Alshehri, Fakhri Karray, Abdulmotaleb El Saddik, Mansour Alsulaiman, and Tiago H Falk. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Information Fusion*, 76:355–375, 2021.
- [107] Apoorva Muley, Prathamesh Muzumdar, George Kurian, and Ganga Prasad Basyal. Risk of ai in healthcare: A comprehensive literature review and study framework. *arXiv preprint arXiv:2309.14530*, 2023.
- [108] Pavankumar Mulgund, Banashri Pavankumar Mulgund, Raj Sharman, and Raghvendra Singh. The implications of the california consumer privacy act (ccpa) on healthcare organizations: Lessons learned from early compliance experiences. *Health Policy and Technology*, 10(3):100543, 2021.
- [109] Alhassan Mumuni and Fuseini Mumuni. Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258, 2022.
- [110] Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, Ghulam Murtaza, Saad Zafar, and Ambreen Bano. Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48:100546, 2023.
- [111] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020.

- [112] Andrew Ng and Michael Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 2001.
- [113] Pablo A. Osorio-Marulanda, Gorka Epelde, Mikel Hernandez, Imanol Isasa, Nicolas Moreno Reyes, and Andoni Beristain Iraola. Privacy mechanisms and evaluation metrics for synthetic data generation: A systematic review. *IEEE Access*, 12:88048–88074, 2024.
- [114] Leandro Pardo. *Statistical inference based on divergence measures*. Chapman and Hall/CRC, 2018.
- [115] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*, 2018.
- [116] Shreyas Patel, Ashutosh Kakadiya, Maitrey Mehta, Raj Derasari, Rahul Patel, and Ratnik Gandhi. Correlated discrete data generation using adversarial training. *arXiv preprint arXiv:1804.00925*, 2018.
- [117] Carl Preiksaitis and Christian Rose. Opportunities, challenges, and future directions of generative artificial intelligence in medical education: scoping review. *JMIR medical education*, 9:e48785, 2023.
- [118] Taki Hasan Rafi, Faiza Anan Noor, Tahmid Hussain, and Dong-Kyu Chae. Fairness and privacy preserving in federated learning: A survey. *Information Fusion*, 105:102198, 2024.
- [119] L Ramos and J Subramanyam. Maverick research: Forget about your real data—synthetic data is the future of ai. *Gartner, Inc, Jun*, 2021.
- [120] Neeta Rana and Hitesh Marwaha. Role of federated learning in healthcare systems: A survey. *Mathematical Foundations of Computing*, 7(4):459–484, 2024.
- [121] Debbie Rankin, Michaela Black, Raymond Bond, Jonathan Wallace, Maurice Mulvenna, Gorka Epelde, et al. Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR medical informatics*, 8(7):e18910, 2020.
- [122] Ifrah Raouf and Manoj Kumar Gupta. A conditional input-based gan for generating spatio-temporal motor imagery electroencephalograph data. *Neural Computing and Applications*, 35(29):21841–21861, 2023.
- [123] Sina Rashidian, Fusheng Wang, Richard Moffitt, Victor Garcia, Anurag Dutt, Wei Chang, Vishwam Pandya, Janos Hajagos, Mary Saltz, and Joel Saltz. Smooth-gan: towards sharp and smooth synthetic ehr data generation. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*, pages 37–48. Springer, 2020.
- [124] Amir Rehman, Huanlai Xing, Li Feng, Mehboob Hussain, Nighat Gulzar, Muhammad Adnan Khan, Abid Hussain, and Dhekra Saeed. Fedcsd-gan: A secure and collab-

- orative framework for clinical cancer diagnosis via optimized federated learning and gan. *Biomedical Signal Processing and Control*, 89:105893, 2024.
- [125] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015.
- [126] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS’18*, page 5234–5243, Red Hook, NY, USA, 2018. Curran Associates Inc.
- [127] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.
- [128] Timur Sattarov, Marco Schreyer, and Damian Borth. Findiff: Diffusion models for financial tabular data generation. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 64–72, 2023.
- [129] Timur Sattarov, Marco Schreyer, and Damian Borth. Fedtabdiff: Federated learning of diffusion probabilistic models for synthetic mixed-type tabular data generation. *arXiv preprint arXiv:2401.06263*, 2024.
- [130] M Scott, Lee Su-In, et al. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30:4765–4774, 2017.
- [131] Thanveer Shaik, Xiaohui Tao, Lin Li, Haoran Xie, and Juan D Velásquez. A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom. *Information Fusion*, 102:102040, 2024.
- [132] Jiayi Shen, Casper JP Zhang, Bangsheng Jiang, Jiebin Chen, Jian Song, Zherui Liu, Zonglin He, Sum Yi Wong, Po-Han Fang, Wai-Kit Ming, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. *JMIR medical informatics*, 7(3):e10010, 2019.
- [133] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
- [134] Loic Simon, Ryan Webster, and Julien Rabin. Revisiting precision recall definition for generative modeling. In *International Conference on Machine Learning*, pages 5799–5808. PMLR, 2019.
- [135] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [136] Mengkai Song, Zhibo Wang, Zhifei Zhang, Yang Song, Qian Wang, Ju Ren, and Hairong Qi. Analyzing user-level privacy attack against federated learning. *IEEE Journal on Selected Areas in Communications*, 38(10):2430–2444, 2020.

- [137] Yige Sun, Jing Li, Yifan Xu, Tingting Zhang, and Xiaofeng Wang. Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, page 120201, 2023.
- [138] Esteban G Tabak and Cristina V Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66(2):145–164, 2013.
- [139] Syed Mahir Tazwar, Max Knobbout, Enrique Hortal Quesada, and Mirela Popa. Tabvae: A novel vae for generating synthetic tabular data. In *ICPRAM*, pages 17–26, 2024.
- [140] Raginee R Titir and Murali Ramanathan. Variational autoencoders for generative modeling of drug dosing determinants in renal, hepatic, metabolic, and cardiac disease states. *Clinical and Translational Science*, 17(7):e13872, 2024.
- [141] Rob Toews. Forbes. Synthetic data is about to transform artificial intelligence. <https://www.forbes.com/sites/robtoews/2022/06/12/synthetic-data-is-about-to-transform-artificial-intelligence/?sh=7d65c55d7523>, 2023. Access Jun-20-2023.
- [142] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. Decaf: Generating fair synthetic data using causally-aware generative networks, 2021.
- [143] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.
- [144] L Vivek Harsha Vardhan and Stanley Kok. Synthetic tabular data generation with oblivious variational autoencoders: alleviating the paucity of personal tabular data for open research. In *Proceedings of the 37th International conference on machine learning, ICML HSYS Workshop 2020*, 2020.
- [145] Alex X Wang, Stefanka S Chukova, Colin R Simpson, and Binh P Nguyen. Challenges and opportunities of generative models on tabular data. *Applied Soft Computing*, page 112223, 2024.
- [146] Lu Wang, Wei Zhang, and Xiaofeng He. Continuous patient-centric sequence generation via sequentially coupled adversarial learning. In *Database Systems for Advanced Applications: 24th International Conference, DASFAA 2019, Chiang Mai, Thailand, April 22–25, 2019, Proceedings, Part II 24*, pages 36–52. Springer, 2019.
- [147] Zhenchen Wang, Barbara Draghi, Ylenia Rotalinti, Darren Lunn, and Puja Myles. High-fidelity synthetic data applications for data augmentation. In Manuel Domínguez-Morales, Javier Civit-Masot, Luis Muñoz-Saavedra, and Robertas Damaševičius, editors, *Deep Learning*, chapter 7. IntechOpen, Rijeka, 2024.
- [148] Kang Wei, Jun Li, Chuan Ma, Ming Ding, Sha Wei, Fan Wu, Guihai Chen, and Thilina Ranbaduge. Vertical federated learning: Challenges, methodologies and experiments. *arXiv preprint arXiv:2202.04309*, 2022.
- [149] Lisa Weijler, Florian Kowarsch, Matthias Wödlinger, Michael Reiter, Margarita Maurer-Granofszky, Angela Schumich, and Michael N Dworzak. Umap based anomaly detection for minimal residual disease quantification within acute myeloid leukemia. *Cancers*, 14(4):898, 2022.

- [150] Mi-Ja Woo, Jerome P Reiter, Anna Oganian, and Alan F Karr. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, 1(1), 2009.
- [151] Jing Wu, Munawar Hayat, Mingyi Zhou, and Mehrtash Harandi. Defense against privacy leakage in federated learning. *arXiv preprint arXiv:2209.05724*, 2022.
- [152] Boming Xia, Qinghua Lu, Liming Zhu, Sung Une Lee, Yue Liu, and Zhenchang Xing. Towards a responsible ai metrics catalogue: A collection of metrics for ai accountability. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pages 100–111, 2024.
- [153] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575, 2018.
- [154] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.
- [155] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, 416:244–255, 2020.
- [156] Mengmeng Yang, Taolin Guo, Tianqing Zhu, Ivan Tjuawinata, Jun Zhao, and Kwok-Yan Lam. Local differential privacy and its applications: A comprehensive survey. *Computer Standards & Interfaces*, page 103827, 2023.
- [157] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. Federated transfer learning. In *Federated Learning*, pages 83–93. Springer, 2020.
- [158] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. Horizontal federated learning. In *Federated Learning*, pages 49–67. Springer, 2020.
- [159] Zeyu Yang, Peikun Guo, Khadija Zanna, and Akane Sano. Balanced mixed-type tabular data synthesis with diffusion models. *arXiv preprint arXiv:2404.08254*, 2024.
- [160] Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, et al. Ehr-safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ Digital Medicine*, 6(1):141, 2023.
- [161] Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9):1–36, 2024.