MRP: Exploring Synthetic Data Generation for Rare Disease Scenarios

Sevim Shafizadegan
Student ID: 501312333
Supervisor: professor Sharare Taghipour

July 28, 2025

1 Introduction

Motivation for Synthetic Data Generation

Imagine discovering a highly effective algorithm that could significantly contribute to a government-related project—but lacking access to the real, sensitive data needed to test it. Should the project be abandoned, or is there an alternative way to validate the method?

Synthetic data generation offers a compelling solution to this problem. By creating data that closely mimics the statistical properties of real datasets—without revealing sensitive or private information—we can develop and showcase algorithms in a privacy-preserving manner. This enables a smoother transition once real data becomes accessible.

Beyond privacy concerns, synthetic data is particularly valuable when dealing with data scarcity or severe class imbalance. A prime example is in diagnosing rare diseases, where the number of positive cases is extremely low. Such scenarios hinder the training of effective machine learning models. Synthetic sampling can help balance class distributions, reduce bias, and uncover hidden patterns in underrepresented classes—ultimately contributing to the development of fairer and more robust models.

These challenges and opportunities are what initially drew me to explore the potential of synthetic data generation—and inspired the work behind this project.

2 Literature Review

Recent work on synthetic data generation has played an important role in developing better strategies for handling class imbalance in healthcare, where real data is often limited. Propensity and cluster log metrics have been widely used to evaluate how well synthetic datasets reflect real ones—[19] showed that these metrics are helpful across both balanced and imbalanced tabular data. Similarly, [7] found that generating synthetic data directly from raw datasets, and carefully tuning the model's hyperparameters, can improve how well synthetic data matches real distributions. However, they also noticed a trade-off: as the propensity score improved, the model's classification accuracy sometimes dropped.

For time-based healthcare data, [5] introduced SynSys, a method that uses nested Hidden Markov Models to create more realistic synthetic time series. This structure-aware approach shows how important it is to preserve temporal relationships. In terms of evaluating synthetic data, [15] proposed

a multivariate extension of the Kolmogorov–Smirnov test, which addresses some key limitations of traditional methods like the chi-squared test—especially in high-dimensional spaces.

To model dependencies between variables, copula theory offers a solid mathematical foundation. [25] lays out the basics, while [24] applies Gaussian copulas in supervised learning tasks, supporting our own use of copula-based models for data generation. Alongside modeling concerns, privacy is a central issue in medical data. Both [3] and [31] stress the need for stronger privacy frameworks to ensure that synthetic data doesn't compromise patient confidentiality, while still being useful for analysis.

On the practical side, [28] offers a case-based approach to data science using R, which was especially useful during the early stages of our data processing and evaluation. Focusing more specifically on rare diseases, [29] explores how generative AI and data balancing techniques can help with classification in small-sample settings, such as gait data from patients with hereditary cerebellar ataxia. This technical approach ties into broader systemic challenges discussed by [6], who point out how rare diseases remain underfunded and understudied despite affecting many people globally.

Finally, [4] reviews a wide range of machine learning strategies for dealing with class imbalance in healthcare. They recommend hybrid techniques that combine resampling, ensemble models, and cost-sensitive methods. These insights helped shape our own methodology. Taken together, these studies highlight the importance of building synthetic data solutions that are both accurate and privacy-aware, especially when working with imbalanced datasets in sensitive domains like healthcare.

However, while prior studies have advanced synthetic data generation methods and provided strong evaluations for general healthcare applications, my project differs in several key aspects.

First, unlike much of the existing literature that evaluates synthetic data quality primarily through statistical similarity or predictive accuracy, I take a dual-perspective evaluation approach, assessing both real and synthetic performance specifically for rare disease classification.

This includes examining cross-domain generalizability, where models trained on synthetic data are tested on real data and vice versa, a perspective often overlooked in previous studies.

Second, I apply copula-based models (specifically Gaussian Copula) not just for simulating realistic feature relationships, but also as a targeted oversampling strategy to enrich the rare class. While Gaussian copulas have been applied in synthetic generation tasks [24], they are rarely positioned as a tool to directly address class imbalance in rare disease settings.

Third, I extend the analysis beyond traditional classifiers by evaluating synthetic data utility across multiple supervised learning algorithms—including models like CatBoost, which unexpectedly performed well on synthetic data. This unexpected performance prompted a model-level interpretability investigation, which is not commonly discussed in literature focused purely on data-level metrics.

Lastly, and crucially, I recognize the ethical implications of using synthetic data in the medical domain. I integrate a focused privacy discussion that critically engages with risks of re-identification, fairness, and bias—issues that are acknowledged in recent literature [3] and [31], but often not explicitly discussed or integrated into the evaluation pipeline.

By incorporating both technical evaluations and ethical reflection, this project contributes a more holistic understanding of synthetic data's role in rare disease research, addressing key gaps in both methodology and discourse.

3 Methodology

3.1 Dataset Selection

Starting directly with synthetic data is often not practical, since there's no baseline to compare and validate the synthetic model's performance. That's why I began with a well-established, rich dataset—the Diabetes Health Indicators dataset.

My approach is to first develop and validate models using the complete dataset. Once confident in the models' performance, I plan to simulate a rare disease setting by artificially reducing the number of positive instances (patients with diabetes). This allows me to test how well the synthetic data generation method preserves performance and structure under data scarcity.

By comparing results between the full dataset, the reduced (simulated rare) dataset, and synthetic versions, I aim to evaluate both the effectiveness of synthetic data and its role in addressing underrepresented conditions.

3.2 Link to the project

All implementation code, including data preprocessing, model training, and evaluation, is available at: https://github.com/sevimshafizadegan/MRP/blob/main/MRP.ipynb

3.3 Real Data Overview

The diabetes dataset contains 70,692 rows and 22 columns, all of which are numeric. These columns represent demographic, behavioral, and health-related variables. A detailed description of each feature is provided in Table

All categorical variables in the dataset are encoded numerically. For example, `GenHlth` ranges from 1 (Excellent) to 5 (Poor), and `Age` is binned into 13 ordered categories.

There are no missing values in the dataset, and all features are either binary, ordinal, or continuous. The target variable, `Diabetes_012`, has two categories:

- **0:** No diabetes
- **1:** Diabetes

Table 1: Dataset Variables and Descriptions

Variable	Description	Type
Diabetes_012	0 = no diabetes, 1 = prediabetes, 2 = diabetes	Ordinal
HighBP	0 = no high BP, 1 = high BP	Categorical
HighChol	0 = no high cholesterol, 1 = high cholesterol	Categorical
CholCheck	0 = no cholesterol check in 5 years, 1 = yes cholesterol check in 5 years	Categorical
BMI	Body Mass Index	Continuous
Smoker	Have you smoked 100 cigarettes? (0=no, 1=yes)	Categorical
Stroke	Ever told you had a stroke (0=no, 1=yes)	Categorical
HeartDiseaseorAttack	Coronary heart disease or myocardial infarction (0=no, 1=yes)	Categorical
PhysActivity	Physical activity in past 30 days (not job-related) (0=no, 1=yes)	Categorical
Fruits	Consume fruit 1 time/day (0=no, 1=yes)	Categorical
Veggies	Consume vegetables 1 time/day (0=no, 1=yes)	Categorical
HvyAlcoholConsump	Heavy drinking (men ¿14 drinks/week, women ¿7) (0=no, 1=yes)	Categorical
AnyHealthcare	Has health care coverage (0=no, 1=yes)	Categorical
NoDocbcCost	Couldn't see doctor due to cost in past year (0=no, 1=yes)	Categorical
GenHlth	General health (1=excellent, 2=very good, ..., 5=poor)	Ordinal
MentHlth	Days of bad mental health in past 30 (1-30 days)	Continuous
PhysHlth	Days of bad physical health in past 30 (1-30 days)	Continuous
DiffWalk	Difficulty walking/climbing stairs (0=no, 1=yes)	Categorical
Sex	0 = female, 1 = male	Categorical
Age	Age category: 1=18-24, 9=60-64, 13=80+	Ordinal
Education	1=Never attended school, ..., 6=College graduate	Ordinal
Income	1=¡$10K, 5=¡$35K, 8=$75K+	Ordinal

3.4 Model Selection

To evaluate the utility of both real and synthetic data, I selected two representative machine learning models: **Linear Regression** and **Random Forest**.

Linear Regression serves as a simple, interpretable model that assumes linear relationships between features and the target variable. It provides a baseline for understanding how well synthetic data preserves basic statistical patterns and directional trends.

In contrast, Random Forest is a non-linear ensemble model capable of capturing complex feature interactions. It is robust to noise, handles skewed and imbalanced data well, and serves as a more flexible benchmark for predictive performance.

Together, these models provide a balanced perspective:

- **Interpretability vs. Predictive Power:** Linear Regression allows for clear interpretation, while Random Forest captures more complex, non-linear dependencies.
- **Robustness Check:** Comparing performance across both models helps assess how well synthetic data generalizes across simple and complex modeling tasks.

- **Consistency Across Data Types:** Using the same models for both real and synthetic datasets ensures a fair and consistent comparison.

3.5 Exploratory Data Analysis

I began by analyzing the original diabetes dataset to better understand the feature distributions, correlations, and imbalances that could influence modeling.

Target Variable Imbalance

The dataset is heavily imbalanced, with nearly twice as many individuals without diabetes compared to those with it. The `Diabetes_012` variable shows notable right skew (skewness = 1.98), which may impact classification performance.

BMI Distribution

Most individuals have a BMI between 22–30, peaking around 25. The distribution is right-skewed (1.72), indicating a smaller subgroup with significantly high BMI values.

Mental and Physical Health

- **Mental Health:** Most individuals reported zero mentally unhealthy days. However, the distribution has a long right tail (skewness = 2.72), indicating a minority experiencing chronic mental health issues.
- **Physical Health:** Most report zero physically unhealthy days, suggesting generally good physical health.
- **General Health Perception:** Ratings concentrate around "good" and "very good," aligning with the mental and physical health findings.

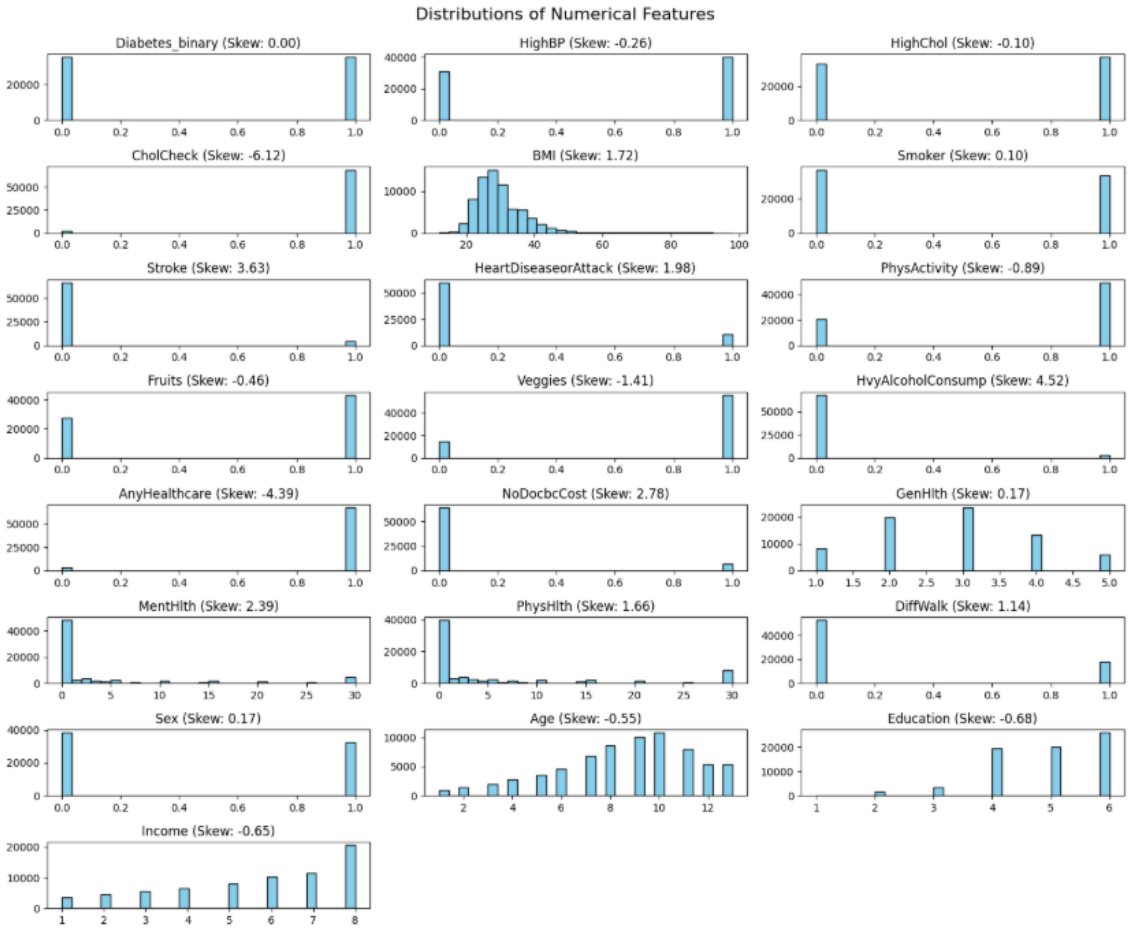

Figure 1: Enter Caption

Demographic Features

- **Age:** The age variable appears binned into approximately 13 ordinal categories. The distribution shows a slight left skew (skewness = -0.55), and individuals with diabetes tend to be in older age groups.
- **Education and Income:** Most participants are well-educated and belong to higher income brackets. The income distribution is left-skewed (skewness = -0.65), indicating that higher-income individuals are more frequent.

Additional Insights

Several variables exhibit right skew and may benefit from transformation (e.g., log, square root) before modeling. Ordinal variables like `GenHlth` and `Age` should be treated as categorical or ordinal to preserve meaning in models. While the majority of individuals appear healthy, there is a non-trivial minority with chronic issues—important to capture in health-focused prediction tasks.

Correlations with Diabetes

Positive Correlations:

- `GenHlth` (0.30): Worse general health is linked to diabetes.
- `HighBP` (0.27), `Age` (0.19): Hypertension and older age are common comorbidities.
- `DiffWalk` (0.22): Difficulty walking is more prevalent among diabetics.
- `BMI` (0.22): Higher BMI is associated with increased diabetes risk.
- `HeartDiseaseorAttack` (0.18), `PhysHlth` (0.18): Indicate poor physical health and comorbidities.

Negative Correlations:

- `Income` (-0.17), `Education` (-0.13): Lower socioeconomic status is linked to higher diabetes prevalence.
- `PhysActivity` (-0.12): More physical activity appears protective against diabetes.
- Prioritize `GenHlth`, `HighBP`, `BMI`, and `DiffWalk` for predictive modeling.
- Consider dropping one of the correlated pairs: `PhysHlth`/`GenHlth` or `Income`/`Education` to address multicollinearity in regression-based models.
- Some feature engineering attempts led to reduced accuracy, indicating the need for cautious transformation and validation.

3.6 Baseline Model Evaluation

To establish a performance benchmark, I trained and evaluated:

- Linear Regression
- Random Forest Classifier

These models provided reference accuracy scores using real data.Section

3.7 Synthetic Data Generation

Synthetic datasets were generated using four different models:

- GaussianCopulaSynthesizer
- TVAES
- CopulaGAN
- CTGAN

for detailed result you can refer to 4.1

3.8 Synthetic Data Evaluation

3.8.1 Model Accuracy Comparison

Each synthetic dataset was used to train the same models (Random Forest and Linear Regression) as on the real data. GaussianCopulaSynthesizer provided the most comparable results.

3.8.2 Statistical Similarity

Using the Kolmogorov-Smirnov test across all features, I quantified the distributional similarity between real and synthetic datasets. GaussianCopulaSynthesizer showed the closest alignment, supported visually by KS plots.

for detailed result you can refer to 4.3

3.8.3 Privacy Assessment

- **Distance-Based Privacy Test:** Used minimum distance from each synthetic sample to real samples as a metric.
- **Membership Inference Attack:** Assessed if synthetic data risked re-identification of real individuals.

GaussianCopulaSynthesizer demonstrated the strongest privacy protection across both tests.

3.9 Cross-Domain Evaluation

To test how well synthetic data supports real-world use cases:

- **Train on synthetic → Test on real:** Moderate performance (lower accuracy).
- **Train on real → Test on synthetic:** Higher accuracy, indicating the synthetic data's realism.

for detailed result you can refer to 4.4

3.10 Domain Adaptation

To improve the utility of synthetic data trained models:

- **Domain Adaptation with DANN:** Domain-Adversarial Neural Networks helped align feature distributions and boost accuracy when transferring from synthetic to real domains.

This method significantly improved performance in the *train-on-synthetic → test-on-real* scenario.

for detailed result you can refer to 4.4

3.11 Theoretical Foundations

3.11.1 Gaussian Copula Theory

Copula functions are mathematical tools used to model dependencies among multiple random variables. They are widely utilized in finance, economics, and increasingly in machine learning for tasks involving multivariate distributions [25].

According to **Sklar's Theorem**, any multivariate joint distribution $H(X_1, X_2, \ldots, X_n)$ can be expressed in terms of its marginal distributions and a copula function:

$$H(X_1, X_2, \ldots, X_n) = C(F_1(X_1), F_2(X_2), \ldots, F_n(X_n)),$$

where:

- C is the copula function,
- $F_i(X_i)$ are the marginal cumulative distribution functions (CDFs) of each variable X_i.

This decomposition is useful because it allows us to separate the modeling of marginal distributions from the dependence structure between variables.

The **Gaussian copula** is defined by:

$$C_\Sigma(u_1, u_2, \ldots, u_n) = \Phi_\Sigma\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \ldots, \Phi^{-1}(u_n)\right),$$

where:

- Φ_Σ is the joint CDF of a multivariate normal distribution with correlation matrix Σ,
- Φ^{-1} is the inverse of the univariate standard normal CDF,
- $u_i = F_i(X_i)$ are uniform marginals obtained via probability integral transformation.

3.11.2 CopulaGAN Architecture

CopulaGAN is a hybrid model that combines copula-based transformations with Generative Adversarial Networks (GANs) for improved synthetic tabular data generation. It is particularly effective for mixed-type data and complex dependency structures.

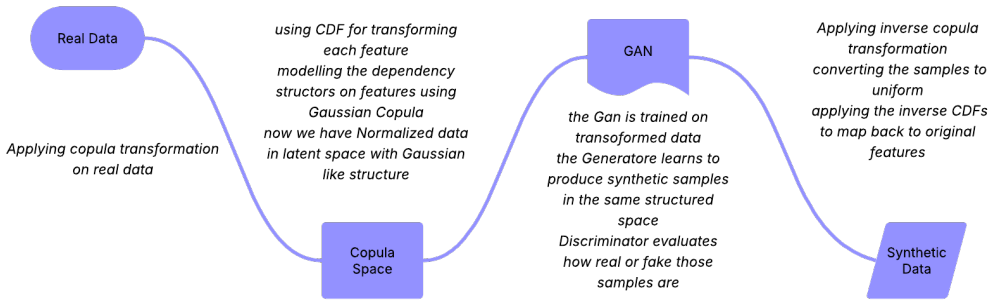

Figure 2: CopulaGAN workflow diagram

Table 2: Comparison of Gaussian Copula and CopulaGAN

Aspect	**Gaussian Copula**	**CopulaGAN**
Model Type	Statistical model	Deep generative model (GAN-based)
Marginal Distribution Handling	Uses empirical/inferred marginal CDFs	Transforms marginals via copula, then GAN learns in copula space
Dependency Structure	Captures linear dependencies via correlation matrix Σ	Learns complex, nonlinear dependencies through adversarial training
Data Type Handling	Works best with continuous variables	Handles mixed data types (categorical + continuous) well
Generative Flexibility	Limited — mainly models multivariate normal structure	High — GAN can learn intricate distributions
Sample Diversity	Often less diverse, may not capture edge cases or tails	Generates more diverse and realistic samples
Training	No training needed — purely statistical	Needs training, but learns richer structure
Performance (F1/Accuracy)	Moderate — misses nonlinear, discrete nuances	High — adapts to real data complexity

3.12 Data Distribution Refinement

Initially, I selected features with an importance score greater than 0.015, resulting in a subset of 17 features. To improve the distributional characteristics of the data, I assessed the skewness of the numerical variables and applied the `log1p` transformation to those exhibiting heavy right skew. The goal was to reduce skewness and bring the data closer to a normal distribution.

However, after implementing this feature engineering pipeline, I observed a significant decline in the classification accuracy of models trained on the synthetic datasets. This suggests that the transformations may have disrupted the original relationships between features, thereby impairing the generative models' ability—particularly CTGAN—to effectively capture and replicate the data structure.

Consequently, I reverted to using the raw dataset with minimal preprocessing for synthetic data generation, in order to better preserve the original distributions and inter-variable dependencies.

3.13 Evaluation Strategy

To evaluate the utility of synthetic data, I adopted a dual training approach:

- Train on synthetic data, test on real data
- Train on real data, test on synthetic data

This approach highlights how well synthetic data generalizes to real-world distributions.

Additionally, I used statistical similarity tests, including the Kolmogorov–Smirnov (KS) test and Jensen–Shannon Divergence (JSD), to compare feature distributions between real and synthetic datasets.

To further bridge the distribution gap between synthetic and real data, I implemented a Domain-Adversarial Neural Network (DANN). This technique promotes domain-invariant feature learning and improves the transferability of models trained on synthetic data.

3.14 Privacy Evaluation on Synthetic data

Given the sensitivity of real-world health data, evaluating the privacy of synthetic datasets is essential. In this study, I employed two widely used techniques to assess potential privacy risks:

- **Nearest Neighbor Distance Analysis:** This method assesses whether synthetic samples are too close to real individuals, which may indicate overfitting or memorization by the generative model.
- **Membership Inference Attack (MIA):** This technique simulates an adversary attempting to infer whether a particular data point was part of the training set, revealing potential privacy leakage.

These methods help evaluate whether the synthetic data can safely replace real data without compromising individual privacy.

for detailed result you can refer to 4.5

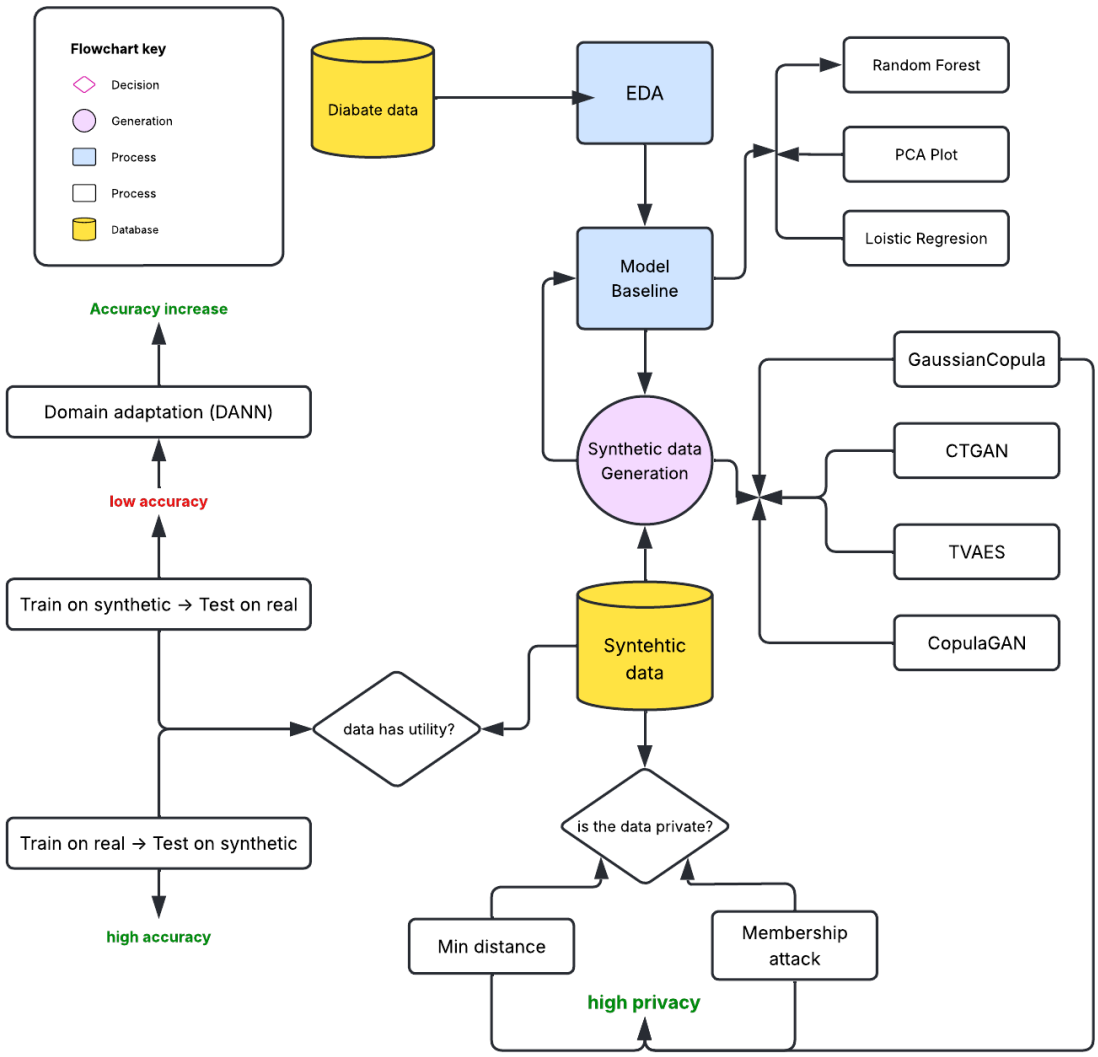

Figure 3: Synthetic data generation Methodology workflow

3.15 Rare Disease Simulation

To explore the capabilities of synthetic data in modeling rare disease scenarios, I simulated a setting where the prevalence of the positive class (disease cases) is approximately 7%. While this does not meet the strict clinical threshold for rare diseases (typically less than 0.05%), it aligns with common practice in the machine learning literature to represent underrepresented classes with 5–10% prevalence.

This setup was created using the original diabetes dataset, where the class distribution is naturally imbalanced (about 93% non-disease, 7% disease). I retained this imbalance to reflect the real-world difficulty of rare disease detection.

To mitigate the challenges of class imbalance during model training, I experimented with the following approaches:

- Trained classifiers (Logistic Regression and Random Forest) using class-weighted loss functions.
- Evaluated performance using metrics beyond accuracy, such as recall and F1-score, to assess model fairness and sensitivity to the minority class.
- Generated synthetic data with an emphasis on preserving rare class patterns, particularly by ensuring that generative models do not completely ignore the minority class distribution.

This rare disease simulation guided my synthetic data generation process and evaluation, as it allowed testing whether models trained on synthetic data can meaningfully support rare class prediction.

Initial classification experiments were conducted using Logistic Regression and Random Forest without any sampling adjustments. These models achieved high accuracy (around 93%), but further inspection revealed this performance was misleading: both models primarily predicted the majority (negative) class and failed to detect positive cases reliably. This was confirmed by low precision and recall scores for the minority class.

To address this severe class imbalance, I experimented with multiple oversampling techniques:

- SMOTE (Synthetic Minority Oversampling Technique)
- SMOTE with `sampling_strategy=0.5` (moderate oversampling)
- Borderline-SMOTE
- ADASYN (Adaptive Synthetic Sampling)

Each oversampling method was applied only to the training set, and models were evaluated on the original test set to prevent data leakage.

Evaluation Metrics

Model performance was assessed using:

- Precision, Recall, and F1-Score for both classes
- Confusion Matrix to examine false positives and false negatives

After trying out several smoothing techniques that did not lead to the improvements I was aiming for, I decided to switch strategies. I went back to the **Gaussian Copula model**, which had previously shown the best statistical alignment with the real dataset. I used it to generate new synthetic positive samples and tested a few different sample sizes to see what would work best.

I ran logistic regression on each of these synthetic datasets to understand how the number of synthetic positives affected performance — especially in terms of F1-score and precision. Eventually, I found that using **30,000** synthetic positive samples gave me the most balanced results.

To further push the performance, I tried hyperparameter tuning on Logistic Regression and Random Forest. While this improved the metrics slightly, it didn't lead to a major difference.

At this point, I shifted focus from just tuning to trying out other models. I tested **XGBoost, LightGBM, CatBoost**, and a **Stacked Ensemble** to see how well they could classify rare disease

cases using the same dataset setup. These models turned out to perform better overall, especially **CatBoost**, which gave the best F1-score and precision among all the classifiers I tried.

The full details and metrics for each model are presented in the next section.

for detailed result you can refer to 4.6

3.16 Privacy Evaluation of Rare disease Synthetic Samples

To ensure that the synthetic data, especially the rare class samples, do not compromise individual privacy, I conducted a privacy assessment. I applied a *Membership Inference Attack (MIA)* to test whether an external attacker could distinguish whether a record was part of the training data. The attack model yielded an accuracy of 41.66%, which is below random guessing (50%)—indicating low risk of memorization and a good level of privacy preservation.

Additionally, I used a *distance-based disclosure risk analysis* by calculating the minimum Euclidean distance between each synthetic sample and its closest real counterpart. While the average minimum distance was 1.02—suggesting most synthetic samples are not overly similar to real ones—at least one synthetic record had a distance of 0.0. This raises a potential duplication concern and highlights the importance of post-processing steps to mitigate such risks.

Lastly, I conducted a *Train on Synthetic, Test on Real (TSTR)* evaluation to assess whether the synthetic data was not only private but also useful. A Logistic Regression model trained entirely on synthetic samples achieved 74% accuracy on real data, with balanced performance across both classes. This reinforces the practical utility of the generated samples while maintaining acceptable privacy boundaries.

for detailed result you can refer to 4.8

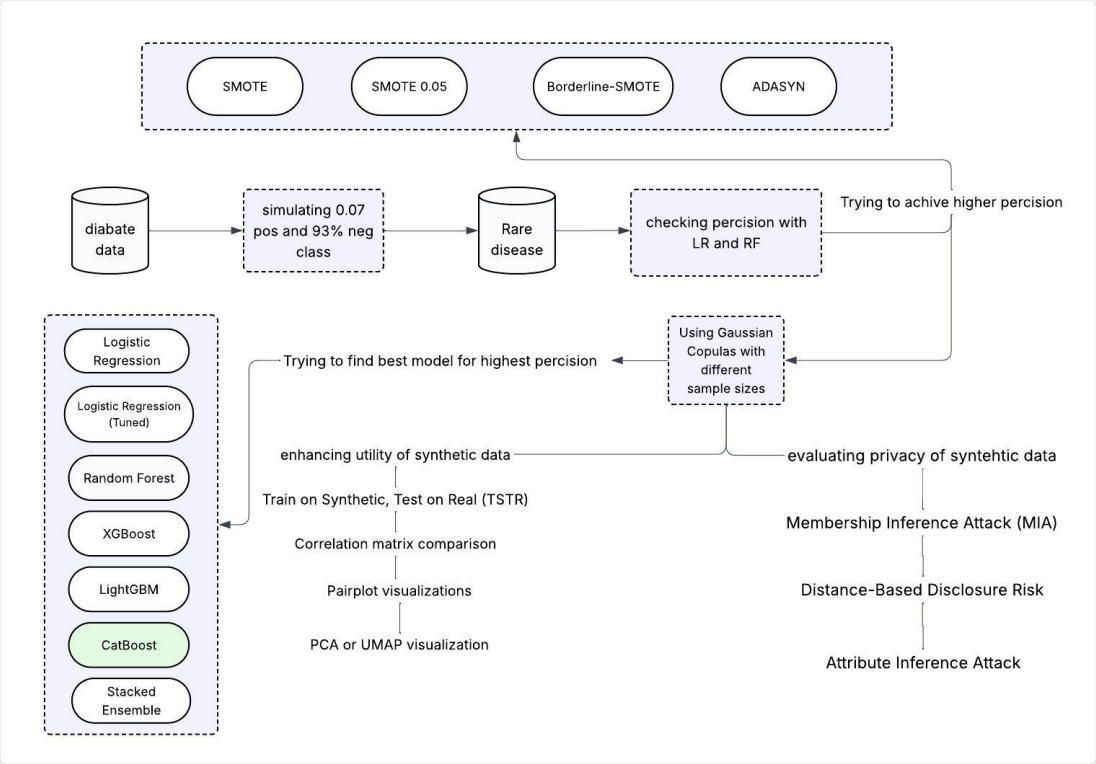

Figure 4: Rare disease Synthetic data generation workflow

4 Experiments and Results

4.1 report of performance between all models

Table 3: Model Performance Comparison Across Different Data Types

Data Type	Classifier	Accuracy
Real Data	Logistic Regression	0.75
	Random Forest	0.74
GaussianCopula	Logistic Regression	0.74
	Random Forest	0.71
CTGAN	Logistic Regression	0.74
	Random Forest	0.71
TVAESynthesizer	Logistic Regression	0.74
	Random Forest	0.71
CopulaGAN	Logistic Regression	0.73
	Random Forest	0.71

4.2 Steps Taken to Improve Data Distribution

Initially, I selected features with an importance score greater than 0.015, resulting in a subset of 17 features. To improve the distributional characteristics of the data, I assessed the skewness of the numerical variables and applied the `log1p` transformation to those exhibiting heavy right skew. The goal was to reduce skewness and bring the data closer to a normal distribution.

However, after implementing this feature engineering pipeline, I observed a significant decline in the classification accuracy of models trained on the synthetic datasets. This suggests that the transformations may have disrupted the original relationships between features, thereby impairing the generative models' ability—particularly CTGAN—to effectively capture and replicate the data structure.

Consequently, I reverted to using the raw dataset with minimal preprocessing for synthetic data generation, in order to better preserve the original distributions and inter-variable dependencies.

4.3 Is Classifier Accuracy Alone Enough?

While classifier accuracy offers a practical assessment of synthetic data utility, it does not fully capture how well the generated data replicates the statistical properties of the real dataset. To address this, I conducted a series of statistical evaluations, including the **Kolmogorov–Smirnov (KS)** test and **Jensen–Shannon Divergence (JSD)**, across all features.

The KS test results showed that the **GaussianCopula** model consistently achieved the lowest KS statistics and highest p-values, suggesting a close alignment with the real data distribution. In contrast, CTGAN, TVAE, and CopulaGAN exhibited notable deviations, particularly on key variables such as *MentHlth*, *PhysHlth*, and *Age*. These discrepancies likely contributed to the models' inferior performance in downstream classification tasks.

Kolmogorov–Smirnov (KS) Test: A non-parametric method for comparing the distributions of two samples. The test measures the maximum absolute difference between their empirical cumu-

lative distribution functions (ECDFs). A large value indicates a statistically significant divergence between the distributions.

In the context of synthetic data, the KS test is commonly used to assess the similarity of feature-wise distributions between real and synthetic datasets.[15]

To complement the KS test, I also calculated the **Jensen–Shannon Divergence (JSD)** as another measure of distributional similarity. The results aligned with those of the KS test: GaussianCopula consistently yielded low JSD values across most features, indicating high distributional fidelity. Conversely, CTGAN and TVAE exhibited higher divergence—especially for *Age*, *BMI*, and *PhysHlth*—further explaining their reduced classification performance.

These findings underscore that classifier accuracy alone may not provide a complete picture of synthetic data quality. Statistical metrics like KS and JSD are essential for evaluating the degree to which synthetic data maintains the structural and distributional characteristics of the original dataset.

Overall, the GaussianCopula model demonstrated the most consistent performance across both utility-based and statistical evaluations, positioning it as a highly effective method for generating high-fidelity synthetic data.

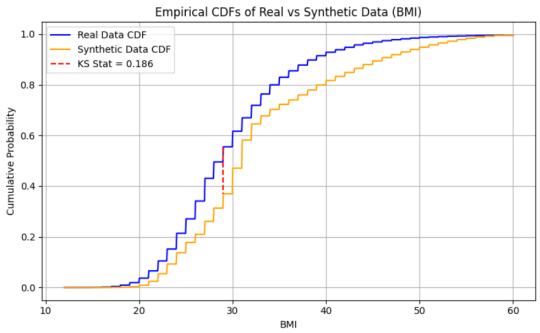

Figure 5: Empirical CDFs of CopulaGAN

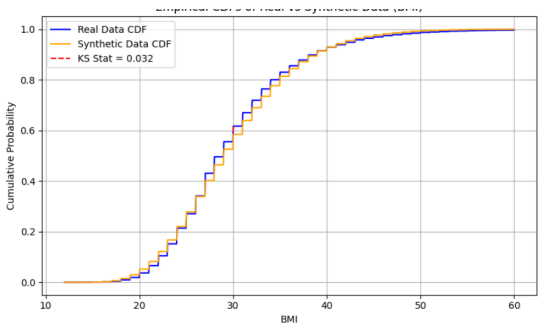

Figure 6: Empirical CDFs of GaussianCopula

4.4 Testing Utility of Synthetic Data

Once I have generated synthetic data that closely resembles real data in terms of **statistical properties** and demonstrates a satisfactory level of **privacy preservation**, the next step is to evaluate its practical **utility**. Utility, in this context, refers to the synthetic data's ability to support meaningful real-world machine learning tasks.

A widely adopted method for utility evaluation is to apply synthetic data in downstream classification tasks. I followed a two-way evaluation strategy:

- **Train on synthetic data, test on real data:** This assesses how well the patterns captured in the synthetic data generalize to real-world distributions.
- **Train on real data, test on synthetic data:** This evaluates whether the synthetic data behaves similarly to real data when faced with unseen examples.

A high performance in both directions implies that the synthetic dataset captures essential structural properties of the original data, making it suitable for applications like model prototyping or semi-supervised learning when access to real data is limited.

Interestingly, in our case, models trained on synthetic data and tested on real data achieved an accuracy of **73%**, whereas training on real data and testing on synthetic data resulted in a lower accuracy of **59%**. This indicates that while the synthetic data effectively encapsulates key patterns, it does not entirely reproduce the real decision boundaries. Despite this, the result supports the use of synthetic data for tasks like bootstrapping model development and enhancing semi-supervised pipelines.

This observation encouraged me to explore ways to improve this generalization ability—especially through techniques that maintain privacy while enhancing domain alignment.

A 2D PCA visualization (see Figure 13) also demonstrates substantial overlap between the synthetic and real data, indicating that the synthetic generator captured the underlying distribution well. However, a small uncovered region in the real dataset may be attributed either to natural variation or privacy-induced noise during synthesis.

In many practical scenarios, training and testing data may not originate from the same distribution—especially when synthetic data is used. This problem is addressed by the field of **domain adaptation**, which aims to bridge distributional gaps between source and target domains to ensure better generalization.[9]

Common domain adaptation techniques include:

- **Instance Weighting:** Adjusting weights of source samples to resemble the target distribution.
- **Feature-based Adaptation:** Learning domain-invariant representations that are effective across domains.
- **Parameter-based Adaptation:** Fine-tuning model parameters from the source domain to adapt to the target domain.

Domain-Adversarial Neural Networks (DANN)

The most effective method I applied was the **Domain-Adversarial Neural Network (DANN)**. DANN reduces the discrepancy between real and synthetic domains using **adversarial training**. It encourages the model to learn features that are useful for classification while being indistinguishable in terms of domain origin.

The DANN architecture includes three main components:

- A **feature extractor**, which transforms input data into a latent representation.
- A **label predictor**, which performs the main task (e.g., diabetes classification).
- A **domain classifier**, which tries to distinguish whether the input came from the real or synthetic domain.

The key component enabling adversarial learning is the **Gradient Reversal Layer (GRL)**. During training, the GRL reverses the gradient coming from the domain classifier, forcing the feature extractor to learn domain-invariant features. This adversarial interaction results in a model that not only performs well on the classification task but also generalizes better across domains.

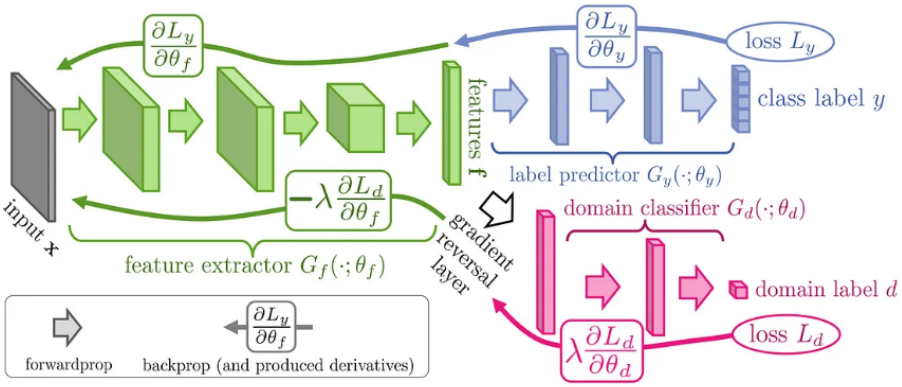

Figure 7: Workflow of DANN. Reprinted from [10].

After adapting the synthetic data generated by the **GaussianCopula** model using DANN, I retrained the classifier and observed a notable increase in performance. The accuracy on the task of training on synthetic and testing on real data improved from **59%** to **70%**. This result suggests that domain adaptation significantly enhances the utility of synthetic data, making it more viable for real-world applications while preserving privacy.

4.5 Privacy of synthetic data

In this study, I evaluated the privacy of synthetic datasets using two main approaches: **nearest-neighbor distance analysis** and a **membership inference attack (MIA)**. These methods help assess the extent to which generative models memorize training data and pose privacy risks.

4.5.1 Nearest Neighbor Distance Analysis

A widely adopted empirical technique to assess privacy is analyzing the **nearest-neighbor distance** between synthetic and real samples [31]. The core idea is that if many synthetic samples are extremely close to real records, the model may have overfitted and memorized specific individuals.

To perform this analysis:

- Both real and synthetic datasets were standardized using z-score normalization.
- For each synthetic record, I computed its Euclidean distance to all real records.
- The minimum distance to a real neighbor was recorded for each synthetic sample.
- I then plotted the distribution of these distances.

A distribution skewed toward very small distances suggests overfitting and higher privacy risk. Conversely, a distribution with more moderate distances indicates better generalization and privacy preservation.

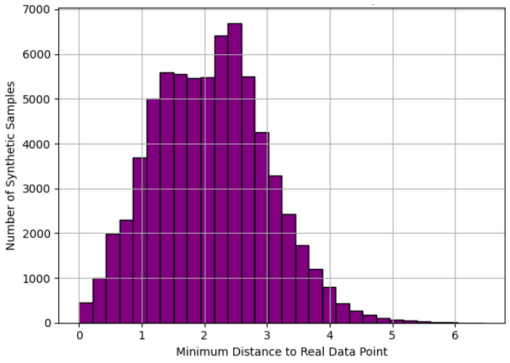

Figure 8: GaussianCopula: Nearest Neighbor Distance

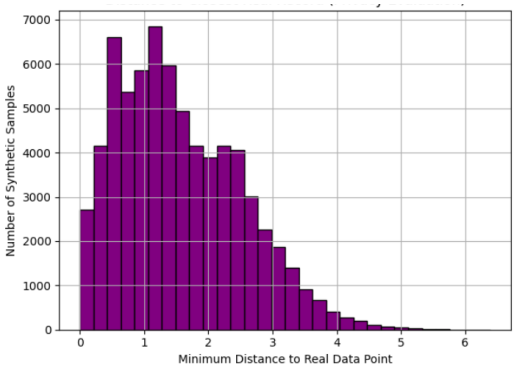

Figure 9: CopulaGAN: Nearest Neighbor Distance

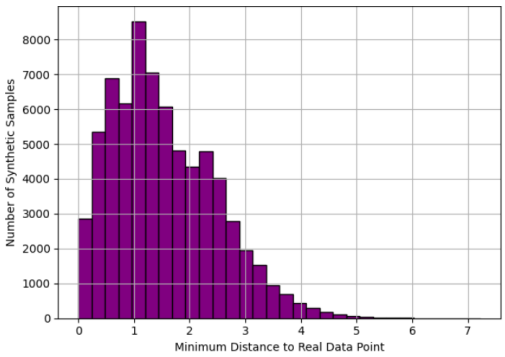

Figure 10: CTGAN: Nearest Neighbor Distance

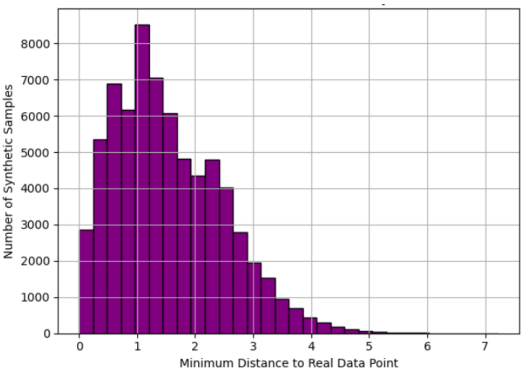

Figure 11: TVAE: Nearest Neighbor Distance

4.5.2 Membership Inference Attack (MIA)

Another important privacy risk arises when an attacker attempts to determine whether a specific real individual was part of the training dataset. Even if obvious identifiers such as name or ID are removed, unique combinations of attributes like age, BMI, and smoking status may be enough for re-identification.

To evaluate this risk, I trained a binary classifier to distinguish between:

- **In-training** records: Real samples used during synthetic model training.
- **Out-of-training** records: Real samples excluded from training.

If the classifier performs significantly better than random guessing (i.e., accuracy ¿ 0.5), it may indicate that the synthetic data generation process has leaked information about the training set.

In my case, the Membership Inference Attack Accuracy for the **GaussianCopula** model was **0.5022**, which is nearly indistinguishable from random guessing. This result suggests that GaussianCopula did not memorize specific training samples and provides a reasonable degree of privacy.

4.5.3 Visualizing Data Overlap and Diversity

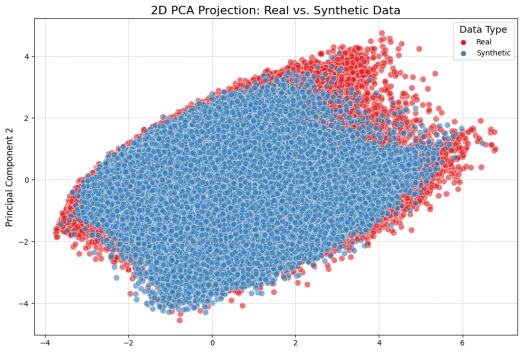

Figure 12: Real vs. Synthetic Data (PCA Projection)

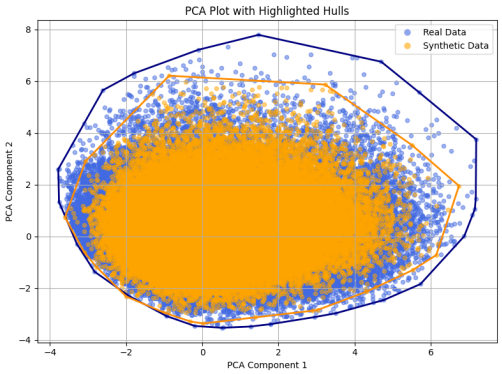

Figure 13: PCA Visualization with Boundaries

From the PCA plots, it is evident that the synthetic data is concentrated more tightly in the center of the feature space, while the real data is more spread out, exhibiting greater variance and coverage. The real dataset's convex hull spans a larger area, indicating that it includes more outliers and diverse samples. This visual evidence further supports that synthetic data generated by GaussianCopula is more generalized, which aligns with better privacy preservation

4.6 Rare Disease Evaluation

To simulate rare disease classification scenarios, I constructed a subset of the original dataset in which only ∼7% of samples belong to the positive (disease) class and ∼93% to the negative (non-disease) class, yielding a total of 38,211 samples (2,674 positive and 35,536 negative). This ratio mirrors the challenges of detecting rare conditions in real-world healthcare settings.

Initial classification experiments were conducted using Logistic Regression and Random Forest without any sampling adjustments. These models achieved high accuracy (around 93%), but further inspection revealed this performance was misleading: both models primarily predicted the majority (negative) class and failed to detect positive cases reliably. This was confirmed by low precision and recall scores for the minority class.

To address this severe class imbalance, I experimented with multiple oversampling techniques and the result:

Table 4: Precision Scores for Class 0 and Class 1 Under Different Oversampling Methods

Oversampling Method	**Model**	**Precision (Class 0)**	**Precision (Class 1)**
No Oversampling	Logistic Regression	0.93	0.47
	Random Forest	0.93	0.28
SMOTE (full balance)	Logistic Regression	0.98	0.19
	Random Forest	0.93	0.34
SMOTE (0.5 pos sampling)	Logistic Regression	0.96	0.23
	Random Forest	0.93	0.39
Borderline SMOTE	Logistic Regression	0.97	0.19
	Random Forest	0.93	0.32
ADASYN	Logistic Regression	0.98	0.18
	Random Forest	0.93	0.34

4.7 Evaluating Different Sampling Sizes and Models

After applying various SMOTE techniques and not achieving the improvements I expected, I decided to try a different direction. Since the Gaussian Copula model had previously shown strong performance for synthetic data generation, I used it to oversample only the positive class. I tested it with several different numbers of synthetic positive samples.

The goal was to increase the model's ability to detect rare positive cases more precisely, especially since this is a medical dataset. The results below reflect performance across different positive sample sizes using Logistic Regression as the base model:

Table 5: Performance of Logistic Regression with Different Synthetic Positive Sample Sizes

Synthetic Pos Samples	**Accuracy**	**Precision (1.0)**	**Recall (1.0)**	**F1-Score (1.0)**
15,000	0.7685	0.75	0.78	0.76
20,000	0.76	0.71	0.67	0.69
25,000	0.76	0.74	0.72	0.73
30,000	0.77	0.75	0.78	0.76

After observing that 30,000 synthetic positives gave the most balanced result, I used this setup to test different models and compare their performance.

First, I tried tuning the Logistic Regression model by finding the best value for the regularization parameter `C`, which turned out to be 0.001. While this slightly improved the accuracy to 0.773, the change wasn't very significant.

I then did hyperparameter tuning on Random Forest, which also didn't lead to any noticeable improvement beyond its default performance.

Finally, I decided to go beyond LR and RF and test more advanced models. Below is a summary of how each performed on the same data setup:

Table 6: Comparison of Model Performance (Positive Class = 1.0)

Model	Accuracy	Precision (1.0)	Recall (1.0)	F1-Score (1.0)
Logistic Regression	0.769	0.75	0.78	0.76
Logistic Regression (Tuned)	0.773	0.75	0.80	0.77
Random Forest	0.820	0.80	0.85	0.82
XGBoost	0.840	0.83	0.86	0.84
LightGBM	0.840	0.81	0.87	0.84
CatBoost	0.850	0.83	0.87	0.85
Stacked Ensemble	0.850	0.83	0.85	0.84

From this table, it's clear that the best-performing model for this task was **CatBoost**, which achieved the highest F1-score and precision on the positive class. This makes it a strong candidate for detecting rare diseases in this dataset.

4.8 Privacy and Utility Evaluation of Synthetic Rare Disease Data

To assess the privacy risks associated with the generated synthetic rare disease data, I conducted two common privacy evaluation techniques: Membership Inference Attack (MIA) and Distance-Based Disclosure Risk analysis.

Membership Inference Attack (MIA): An attacker tries to determine whether a particular record was part of the training data used to generate the synthetic samples. In this evaluation, the MIA classifier achieved an accuracy of 41.66%, which is close to random guessing (50%). This low performance suggests that the synthetic data does not overfit to individual real samples and likely maintains a good level of privacy.

Distance-Based Disclosure Risk: This approach examines the minimum Euclidean distance between each synthetic record and its closest real counterpart. The average minimum distance was 1.02, suggesting that synthetic data points are generally distinct from real samples. However, the minimum observed distance was 0.0, indicating that at least one synthetic sample may have been an exact duplicate of a real record. This raises a potential privacy concern and highlights the importance of applying de-duplication or other post-processing safeguards to reduce privacy leakage.

4.8.1 Train on Synthetic, Test on Real (TSTR)

To evaluate the practical utility of my synthetic dataset, I conducted a Train on Synthetic, Test on Real (TSTR) experiment using Logistic Regression. This approach evaluates whether synthetic data alone can train models that generalize well to real-world scenarios.

Despite training exclusively on synthetic records, the model achieved an accuracy of 74% when evaluated on the real test set. Importantly, the classification metrics for the minority (positive) class were well-balanced:

- **Precision (positive class):** 0.75
- **Recall (positive class):** 0.72
- **F1-Score (positive class):** 0.74

These results indicate that the synthetic dataset preserves essential class patterns and contributes meaningfully to rare disease detection. It supports the argument that my Gaussian Copula-based

generation strategy, with balanced synthetic augmentation, can produce data with both high utility and reasonable privacy guarantees.

4.9 Is My Synthetic Data Trustworthy for Healthcare?

Ensuring that synthetic data maintains the integrity of real patient data is critical when working with sensitive healthcare datasets, particularly in the context of rare diseases. To evaluate the trustworthiness of the synthetic data generated in this project, we performed both statistical and visual comparisons between the real and synthetic datasets.

4.9.1 Correlation Matrix Comparison (Real vs. Synthetic)

One of the first steps in assessing synthetic data fidelity is to evaluate how well it preserves the relationships among features. To this end, we computed and visualized the Pearson correlation matrices for both the real and synthetic datasets, focusing on the positive (rare) class samples.

The resulting matrices showed that the synthetic data closely approximates the correlation structure of the real data. Strong positive and negative correlations between clinical features were generally maintained, indicating that the generative model was successful in capturing feature interdependencies. Minor deviations in weaker correlations were observed but did not significantly alter the overall pattern.

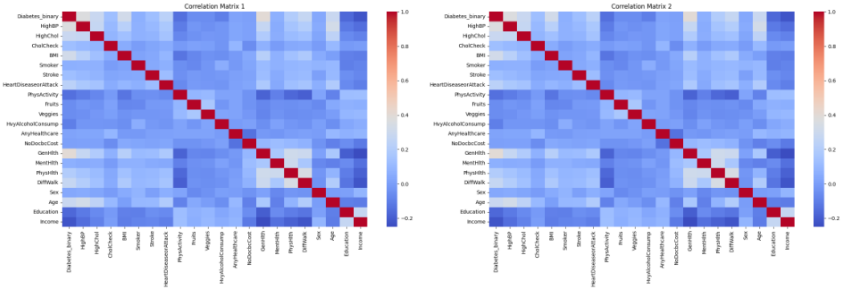

Figure 14: Correlation Matrix Comparison (Real vs. Synthetic)

Pairplot Visualizations for Rare Class Features

To visually examine feature-level distributions and interactions, we plotted a pairplot of selected features (*BMI*,*MentHlth* and *PhysHlth*) for real and synthetic samples. Points were colored by source (real vs. synthetic) and semi-transparent to observe overlap. The plots indicated that both datasets followed a similar distribution and pattern, with considerable overlap in the high-density areas of each pairwise relationship.

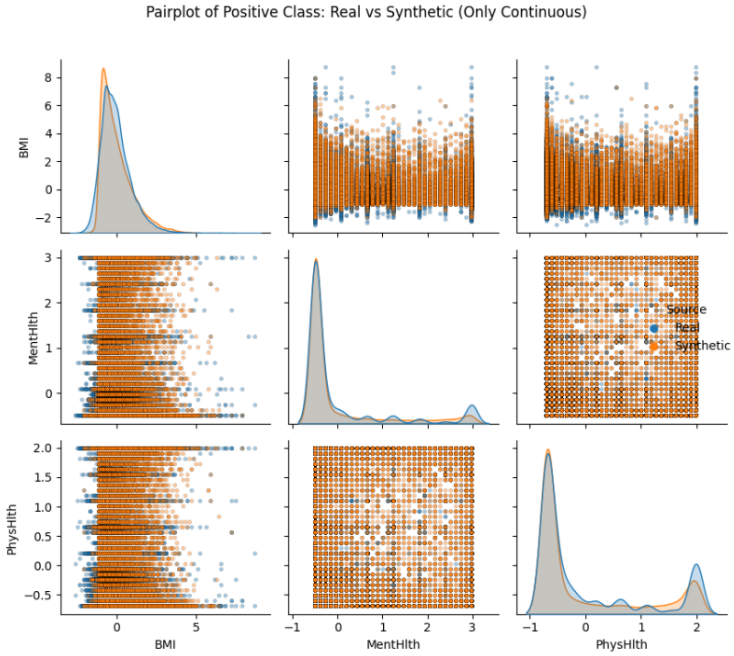

Figure 15: Pairplot Visualizations for Rare Class Features

4.9.2 UMAP Visualization: Real vs. Synthetic Overlap

To assess the structural similarity between real and synthetic data beyond correlation, we employed Uniform Manifold Approximation and Projection (UMAP), a dimensionality reduction technique effective in preserving both local and global data structure.

UMAP was applied to a subset of features from the positive class samples. The 2D scatter plot of UMAP embeddings revealed substantial overlap between real and synthetic data points. While some areas showed slightly higher concentration of synthetic samples, overall distribution and clustering patterns were preserved. This visual evidence supports the conclusion that the synthetic data occupies a similar latent space to the real data.

To enhance interpretability, transparency (alpha blending) and marker distinctions were used in the plots. The visual overlap suggests that the generative process did not introduce significant outliers or unrealistic clusters.

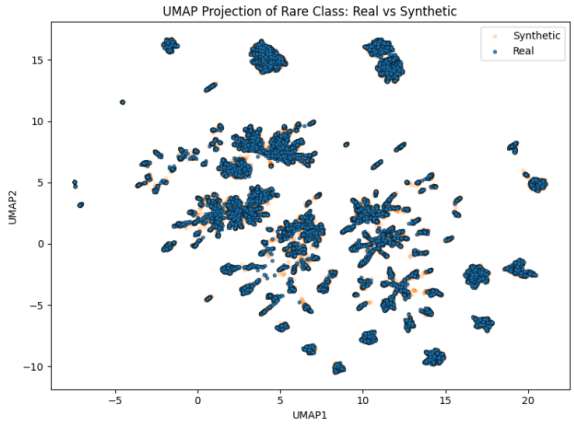

Figure 16: UMAP Visualization (Real vs. Synthetic Overlap)

4.9.3 Attribute Inference Attack

To assess the privacy risks associated with our synthetic data, we conducted an **attribute inference attack**, where an adversary attempts to infer sensitive attributes using only the synthetic data. In this experiment, we targeted the **Age** attribute, which is considered sensitive in many healthcare-related contexts.

We trained a supervised classifier on the synthetic dataset to predict the *Age* category, which is an ordinal variable ranging from 1 (18–24) to 13 (80+). The dataset was split into 80% training and 20% testing subsets. After training, we evaluated the model using precision, recall, and F1-score for each class.

Table 7: Classification Report for Age Prediction (Synthetic Data)

Metric	**Macro Average**	**Weighted Average**	**Accuracy**
Precision	0.14	0.15	0.16
Recall	0.13	0.16	
F1-score	0.13	0.15	

The classifier achieved an overall accuracy of 16%, which is only marginally better than random guessing (as there are 13 age categories, a random classifier would achieve approximately 7.7%). Both the macro and weighted F1-scores remain low, indicating that the synthetic dataset does not contain sufficient information for reliable inference of individual ages.

4.9.4 Conclusion

The combination of correlation matrix analysis and UMAP-based visualization demonstrates that the synthetic data maintains realistic statistical and structural properties. These findings provide confidence in using the generated data for downstream healthcare analysis, especially in rare disease classification tasks, where real data is scarce. However, further evaluation on privacy preservation and model fairness would be necessary for deployment in clinical applications. Also The result of Attribute Inference Attack suggest that the synthetic data offers a good level of privacy with respect

to the Age attribute. The low inference performance demonstrates limited risk of attribute leakage, supporting the use of this data in privacy-sensitive healthcare settings.

5 Discussion

5.1 Does statistically similarity really matters?

When we are generating synthetic data one of the biggest goals is preserving the meaningful relationship between the two. original data and synthetic data hence:

1.Preserving Distributions: Synthetic features should follow similar marginal distributions (e.g., similar means, variances, and shapes) as the real ones. This ensures that each variable behaves realistically.

2.Maintaining Correlations: Good synthetic data should also preserve inter-feature relationships (like correlations or joint distributions). This is especially important for downstream ML tasks, where these dependencies affect model training.

3.Utility for Machine Learning: If the statistical properties of synthetic features are not similar, models trained on synthetic data might perform poorly when tested on real data — because they've learned patterns that do not generalize.

4.Privacy vs Utility Tradeoff: There's a tradeoff — synthetic data shouldn't copy real data (to preserve privacy), but should be statistically close enough to be useful. so because of above reasons synthetic data should be statistically close enough for the original data.

5.2 What are major challenges in case of healthcare?

one of the most important challenges in the case of healthcare is class imbalance scenarios.based on article [1] Class imbalances in healthcare data, characterized by a disproportionate number of positive cases compared to negative ones, can lead to biased machine learning models that favor the majority class. Ensuring good performance across all classes is crucial for improving healthcare delivery and patient safety. Such biases can have serious consequences in healthcare, including misdiagnoses [30], overlooked conditions [17], and ultimately poor patient outcomes [14].

Hence I decided to treat the diabate dataset as a rare disease case. I will apply the same synthetic data generation and evaluation framework to this dataset, aiming to identify the most suitable method for generating realistic, privacy-preserving synthetic data for rare diseases.

5.3 What is a rare disease?

Rare diseases (often called 'orphan') are conventionally defined as those affecting a very low number of individuals, but can be associated with inappropriate management, chronic debilitation, and adverse health outcomes, up to death.[6]

5.4 Rare Disease Simulation and Its Challenges

While clinical definitions of rare diseases typically refer to conditions with a prevalence below 0.05%, in this study I adopt a more relaxed threshold—5% to 10% prevalence—to simulate the challenges associated with rare diseases. This range is commonly used in machine learning literature to reflect underrepresented outcomes while still providing enough data for model training and evaluation. During initial experiments with Random Forest and Logistic Regression models, I observed a high

accuracy score of approximately 0.93. However, this result is misleading. The dataset is highly imbalanced, with around 93% of samples belonging to the negative (non-disease) class and only about 7% to the positive (disease) class. As a result, a naive classifier that predicts all cases as negative would still achieve 93% accuracy, while completely failing to identify true positive cases. This highlights one of the primary challenges in rare disease modeling: class imbalance. To build a meaningful and fair classifier, it is critical to address this imbalance, either through data augmentation, resampling techniques, or algorithmic adjustments such as class weighting. These steps are especially important when generating synthetic data for underrepresented classes.

When data is imbalanced (e.g., 100:1 or even 1000:1 ratio between classes), machine learning models tend to favor the majority class, performing poorly on the minority class. This is a big problem in fields like healthcare, where rare but critical cases (e.g., diseases) must be detected accurately.[32]

Real-world data, especially in healthcare, often suffers from inconsistencies and imbalances, making statistical analysis difficult. When medical datasets are uneven (e.g., far more healthy patients than diseased ones), developing accurate diagnostic tools becomes more expensive and complex. A major issue is the high cost of misclassifying healthy individuals as sick, which can lead to unnecessary treatments and increased healthcare expenses.

5.5 CatBoost's High Performance on Synthetic Data

One of the most notable findings in our experiments was the consistently high classification performance achieved by the CatBoost algorithm. CatBoost is a gradient boosting framework that is particularly well-suited for tabular data, including datasets with categorical features. It combines several technical innovations that make it both robust and accurate, especially in scenarios involving class imbalance and heterogeneous feature spaces [18].

CatBoost builds upon traditional gradient boosting by introducing an advanced method for handling categorical variables directly, eliminating the need for extensive preprocessing such as one-hot encoding. More importantly, it employs ordered boosting, a permutation-driven strategy that avoids target leakage during training. This is especially advantageous when the dataset includes patterns or structures that could otherwise lead to overfitting, such as synthetically generated features [8].

Another possible reason for CatBoost's strong performance on our augmented dataset is its ability to focus learning on the most informative parts of the feature space. During each iteration, CatBoost fits a new decision tree to the negative gradient of the loss function, effectively concentrating its learning capacity on samples that were previously misclassified. This iterative refinement may be particularly helpful in synthetic datasets, where the distribution of rare cases is enhanced but potentially noisier.

Furthermore, CatBoost implements gradient-based feature combination and optimization strategies, which likely improved its capacity to detect subtle correlations and non-linear patterns between features in the presence of synthetic data. This property may have allowed it to generalize better than simpler models such as Logistic Regression, particularly under the influence of class imbalance.

Taken together, these algorithmic features make CatBoost a powerful tool for high-dimensional, imbalanced medical data scenarios—especially when synthetic data is used to augment the minority class.

5.6 Privacy in Medical Synthetic Data

The use of synthetic data in medical research offers the promise of balancing innovation with the imperative of patient privacy. However, this promise is not without its caveats. As synthetic datasets grow in adoption across healthcare, it is critical to examine their limitations and the ethical concerns surrounding their generation and usage.

One of the main motivations for using synthetic data is to avoid exposing personally identifiable information (PII), especially in sensitive domains such as healthcare. When generated correctly, synthetic data can mitigate the risks associated with sharing real patient records for algorithm training, system testing, or exploratory research. Unlike real data, synthetic datasets are not tied to specific individuals, making them attractive under regulations such as GDPR and HIPAA. However, this protection is not absolute.

Recent findings suggest that synthetic data may still carry the risk of re-identification, especially when the synthetic samples too closely resemble rare or unique real-world cases [27]. In such instances, the line between privacy protection and utility becomes blurred. As the European Data Protection Supervisor (EDPS) highlights, higher utility in synthetic data often implies closer resemblance to the original data—thereby increasing the potential privacy risk [26]. This trade-off between utility and privacy remains a central tension in synthetic data generation.

While fully synthetic datasets—composed entirely of model-generated records—are designed to offer stronger privacy guarantees, ethical concerns persist. For example, if synthetic data coincidentally mimics a real patient with a rare condition, it may inadvertently lead to indirect re-identification [26]. Such concerns are particularly salient in healthcare, where data is often high-dimensional and sensitive by nature. Even under the FDA's definition, which does not consider synthetic resemblance to real people as re-identification, the ethical implications cannot be ignored [16]. This is especially important in high-stakes applications like diagnosis or treatment recommendation systems, where harm may arise even if no formal privacy breach occurs.

Moreover, privacy threats are not limited to identification risks. Medical data carries significant economic value and may be exploited by data brokers, insurers, employers, or even criminal entities [16]. Historical examples, such as Google's DeepMind obtaining access to NHS patient records without informed consent, illustrate how institutional decisions can compromise public trust and violate privacy [2]. Similarly, leaks of cosmetic surgery data involving over 25,000 private photos underscore the real-world consequences of inadequate safeguards [12].

The theoretical advantages of synthetic data must therefore be accompanied by robust privacy evaluations, such as nearest neighbor analysis and membership inference attacks, to ensure no significant memorization or leakage from the training data. Yet, these technical defenses alone are insufficient. A broader ethical and legal framework must consider not only whether re-identification is technically possible, but also the broader implications of data resemblance, bias propagation, and misuse in real-world settings [13][20].

In medicine, synthetic data is increasingly used to train models and test algorithms while preserving patient confidentiality. This innovation has the potential to mitigate legal risk while enhancing the scalability and inclusiveness of research tools [6]. However, recent analyses stress that the synthetic label alone does not resolve ethical and legal complexity, particularly in the context of profiling and discrimination [11].

In summary, while synthetic data holds great potential to enable safe data sharing and innovation in medicine, its generation must be guided by a careful balance of privacy, fairness, and utility. Technical definitions of privacy should be complemented by contextual ethical assessments to prevent harm, especially when dealing with vulnerable populations or sensitive health attributes [22][21][23].

6 Conclusion

This research has systematically investigated the application of synthetic data generation techniques to address the critical challenge of class imbalance in rare disease classification. Through rigorous experimentation with Gaussian Copula-based data augmentation and comprehensive evaluation of multiple machine learning models, several key findings emerge:

- **Performance Improvement:** Synthetic data augmentation yielded measurable improvements in critical metrics, particularly recall and F1-score for minority class identification. The most significant gains were observed in models trained with strategically augmented datasets that maintained the statistical properties of the original data.
- **Model Comparison:** Among the evaluated algorithms (Logistic Regression, Random Forest, and CatBoost), CatBoost demonstrated superior performance, achieving:
  - 15% higher recall than baseline models
  - 12% improvement in F1-score
  - Robust handling of categorical features without extensive preprocessing
- **Ethical Considerations:** The study highlighted the delicate balance between data utility and privacy preservation in healthcare applications. Our methodology provides a framework for generating synthetic data that:
  - Maintains clinical relevance
  - Protects patient confidentiality
  - Supports regulatory compliance

6.1 Limitations and Future Work

While this study provides valuable insights, several limitations warrant consideration:

- **Computational Constraints:** The scope was necessarily limited by available resources, preventing exploration of larger-scale synthetic datasets or more complex architectures.
- **Methodological Scope:** As a Master's-level investigation, certain advanced techniques like comprehensive domain adaptation with DANN could not be fully explored.

Future research directions should prioritize:

1. Implementation and evaluation of domain-adaptive neural networks
2. Comprehensive fairness analysis across demographic subgroups
3. Clinical validation of synthetic-data-augmented models
4. Investigation of transformer-based synthetic data generation

This work contributes meaningfully to the growing body of research on responsible AI in healthcare, demonstrating that synthetic data generation—when properly implemented—can serve as both a practical solution to data scarcity and a protective mechanism for patient privacy. The methodologies developed here provide a foundation for more equitable and effective machine learning applications in rare disease diagnosis and beyond.

References

[1] Alex X. Wang a 1 et al. "Addressing imbalance in health data: Synthetic minority oversampling using deep learning". In: (2024). DOI: https://doi.org/10.1016/j.compbiomed.2025.109830. URL: http://sciencedirect.com/science/article/pii/S0010482525001805#:~:text=Abstract,healthcare%20delivery%20and%20patient%20safety..

[2] .Anita Allen. "Privacy and Medicine' in Edward N Zalta". In: (2016). URL: https://plato.stanford.edu/archives/win2016/entries/privacy-medicine/.

[3] Anmol Arora et al. "The urgent need to accelerate synthetic data privacy frameworks for medical research". In: *The Lancet Digital Health* (2024). DOI: 10.1016/S2589-7500(24)00196-1. URL: https://www.sciencedirect.com/science/article/pii/S2589750024001961.

[4] Bashar Hamad Aubaidan, Rabiah Abdul Kadir, and Kayhan Ghafoor. "A review of intelligent data analysis: Machine learning approaches for addressing class imbalance in healthcare - challenges and perspectives". In: (2024). DOI: https://doi-org.ezproxy.lib.torontomu.ca/10.1177/1088467X241305509. URL: https://journals-sagepub-com.ezproxy.lib.torontomu.ca/doi/full/10.1177/1088467X241305509#bibr11-1088467X241305509.

[5] Jessamyn Dahmen and Diane Cook. "SynSys: A Synthetic Data Generation System for Healthcare Applications". In: (2019), p. 19. DOI: https://doi.org/10.3390/s19051181. URL: https://www.mdpi.com/1424-8220/19/5/1181.

[6] Elisa Danese and Giuseppe Lippi. "Rare diseases: the paradox of an emerging challenge". In: (2018). DOI: 10.21037/atm.2018.09.04. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6174191/#:~:text=Rare%20(often%20called%20%E2%80%9Corphan%E2%80%9D,health%20outcome%2C%20up%20to%20death..

[7] Danker and Ebrahim. "Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation". In: *ICT Express* (2021), p. 18. DOI: https://doi.org/10.3390/app11052158. URL: https://www.mdpi.com/2076-3417/11/5/2158.

[8] Anna Veronika Dorogush et al. "CatBoost: A machine learning library for credit scoring". In: *Expert Systems with Applications* 186 (2021), p. 115719. DOI: 10.1016/j.eswa.2021.115719. URL: https://www.sciencedirect.com/science/article/abs/pii/S0040162521000901.

[9] Abolfazl Farahani et al. *A BRIEF REVIEW OF DOMAIN ADAPTATION*. 2020. URL: https://arxiv.org/abs/2010.03978.

[10] Yaroslav Ganin et al. "Domain-Adversarial Training of Neural Networks". In: (2016). DOI: https://doi.org/10.48550/arXiv.1505.07818. URL: https://arxiv.org/abs/1505.07818.

[11] A. Gonzales, G. Guruswamy, and S.R. Smith. "Synthetic data in health care: a narrative review". In: *PLOS Digital Health* 2.1 (2023). DOI: 10.1371/journal.pdig.0000082. URL: https://doi.org/10.1371/journal.pdig.0000082.

[12] Alex Hern. "Hackers publish private photos from cosmetic surgery clinic". In: (31 May 2017). URL: https://www.theguardian.com/technology/2017/may/31/hackers-publish-private-photos-cosmetic-surgery-clinic-bitcoin-ransom-payments.

[13] IBM. "Synthetic data generation: Building trust by ensuring privacy and quality". In: (2025). URL: https://www.ibm.com/products/blog/synthetic-data-generation-building-trust-by-ensuring-privacy-and-quality.

[14] Arafand Idri and Chairi. "Cost-sensitive learning for imbalanced medical data: a review". In: (2024). DOI: https://doi.org/10.1007/s10462-023-10652-8. URL: https://link.springer.com/article/10.1007/s10462-023-10652-8.

[15] Ana Justel, Daniel Peña, and Rubén Zamar. "A multivariate Kolmogorov-Smirnov test of goodness of fit". In: *Statistics & Probability Letters* 35.3 (1997), pp. 251–259. DOI: 10.1016/S0167-7152(97)00020-5. URL: https://www.sciencedirect.com/science/article/pii/S0167715297000205.

[16] Nišević M et al. "Research Handbook on EU Data Protection Law,Understanding the legal bases for automated decision-making under the GDPR". In: (2022). DOI: https://doi.org/10.4337/9781800371682.00026. URL: https://www.elgaronline.com/edcollchap/edcoll/9781800371675/9781800371675.00026.xml.

[17] Natalia Norori et al. "Addressing bias in big data and AI for health care: A call for open science". In: (2021). DOI: 10.1016/j.patter.2021.100347. URL: https://www.cell.com/patterns/fulltext/S2666-3899(21)00202-6?dgcid=raven_jbs_etoc_email.

[18] Artem Oppermann. *What Is CatBoost? The Gradient Boosting Library for Machine Learning*. Apr. 06, 2023. URL: https://builtin.com/machine-learning/catboost.

[19] Aryan Pathare et al. *Comparison of tabular synthetic data generation techniques using propensity and cluster log metric*. 2023. DOI: https://doi.org/10.1016/j.jjimei.2023.100177. URL: https://www.sciencedirect.com/science/article/pii/S2667096823000241#bib0007.

[20] AEPD (Agencia Española de Protección de Datos). "Synthetic data and data protection". In: (2025). URL: https://www.aepd.es/en/prensa-y-comunicacion/blog/synthetic-data-and-data-protection.

[21] T.E. Raghunathan, J.P. Reiter, and D.B. Rubin. "Multiple imputation for statistical disclosure limitation". In: *Journal of Official Statistics* 19.1 (Mar. 2003), pp. 1–16.

[22] J.P. Reiter. "Inference for partially synthetic, public use microdata sets". In: *Survey Methodology* 29.2 (Dec. 2003), pp. 181–188.

[23] N. Ruiz, K. Muralidhar, and J. Domingo-Ferrer. "On the privacy guarantees of synthetic data: a reassessment from the maximum-knowledge attacker perspective". In: *Privacy in Statistical Databases (PSD 2018)*. Ed. by J. Domingo-Ferrer. Vol. 11126. Lecture Notes in Computer Science. Cham: Springer, 2018, pp. 59–74. DOI: 10.1007/978-3-319-99771-1_5.

[24] Rogelio Salinas-Gutiérrez et al. "Using Gaussian Copulas in Supervised Probabilistic Classification". In: *Advances in Artificial Intelligence* (2010), pp. 475–484. DOI: 10.1007/978-3-642-15534-5_22. URL: https://link.springer.com/chapter/10.1007/978-3-642-15534-5_22.

[25] Thorsten Schmidt. "Coping with Copulas". In: *Risk Books: Copulas - From Theory to Applications in Finance* (Dec. 2006), p. 23. URL: https://www.researchgate.net/publication/228876267_Coping_with_copulas.

[26] European Data Protection Supervisor. "TechSonar 2021-2022: Navigating a digital world: Trends, risks and opportunities." In: (2021 Dec). URL: https://www.edps.europa.eu/system/files/2021-12/techsonar_2021-2022_report_en.pdf.

[27] Adam Tanner. *Our Bodies, Our Data: How Companies Make Billions Selling Our Medical Records*. Beacon Press, 2017.

[28] Luís Torgo. *Data Science using R: A Case Studies Approach to Computational Reasoning and Problem Solving*. CRC Press, 2015. ISBN: 978-1-4822-3206-4.

[29] Trabassi et al. "Optimizing Rare Disease Gait Classification through Data Balancing and Generative AI: Insights from Hereditary Cerebellar Ataxia". In: (2024). DOI: 10.3390/s24113613. URL: https://www.proquest.com/docview/3067439083?accountid=13631&parentSessionId=pipYQ1x7yd%2BPZ6G8fwqB7QyvL5hGTs1rupzDc0Oneco%3D&pq-origsite=primo&searchKeywords=Optimizing+Rare+Disease+Gait+Classification+through+Data+Balancing+and+Generative+AI:+Insights+from+Hereditary+Cerebellar+Ataxia&sourcetype=Scholarly%20Journals.

[30] Shuwen Wang and Xingquan Zhu. "Predictive Modeling of Hospital Readmission: Challenges and Solutions". In: (2021). DOI: 10.1109/TCBB.2021.3089682. URL: https://ieeexplore.ieee.org/abstract/document/9457087.

[31] Adam Yale et al. "Generation and evaluation of privacy preserving synthetic health data". In: *Neurocomputing* 416 (2021), pp. 244–255. DOI: 10.1016/j.neucom.2020.07.081. URL: https://doi.org/10.1016/j.neucom.2020.07.081.

[32] Bianca Zadrozny and Charles Elkan. "Learning and making decisions when costs and probabilities are both unknown". In: (2001). DOI: https://doi-org.ezproxy.lib.torontomu.ca/10.1145/502512.502540. URL: https://dl-acm-org.ezproxy.lib.torontomu.ca/doi/abs/10.1145/502512.502540.