Anonymizing medical data for use in AI modeling

Albert Khaidarov

Mentor: dr. Janez Žibert

FAMNIT, April 2025

Motivation & Background

- Real-world data is limited by privacy, cost, and accessibility.
- Synthetic data offers a privacy-safe alternative.
- Useful in training ML models, testing, simulation.

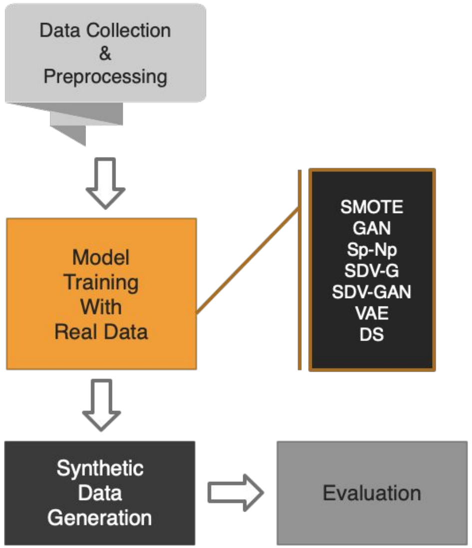
🗣 *“Why generate data? Because using real data is increasingly risky, limited, or expensive.”*

What Is Synthetic Data?

- **Definition:** Artificially created data that mirrors the statistical patterns of real-world datasets.
- **Types:** Tabular, time-series, image, text.
- **Methods:**
  - Statistical simulation
  - Oversampling (e.g., SMOTE)
  - Machine learning models
  - Deep learning (GANs, VAEs)

Reference:

- Jordon et al., 2018; Goodfellow et al., 2014



Architecture of Synthetic data generation. Source: https://dl.acm.org/doi/abs/10.1145/3548785.3548793

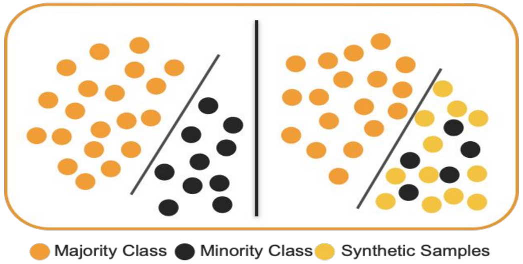
Methods – How Synthetic Data Is Created

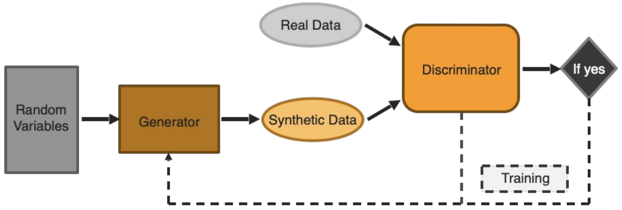
SMOTE (Synthetic Minority Oversampling Technique):

- Generates new samples by interpolating between existing data points in minority class
- Commonly used for handling imbalanced datasets
- Simple, fast, and effective — but limited to basic patterns

GANs (Generative Adversarial Networks):

- Two neural networks (Generator & Discriminator) play a game
- Generator creates fake data; Discriminator tries to spot fakes
- Used for complex, high-quality synthetic data (e.g., tabular, images, time-series)





SMOTE and GANs representations. Source: https://dl.acm.org/doi/abs/10.1145/3548785.3548793

Tools & Technologies

Brief overview of tools:

- **CTGAN** – Tabular GAN
- **DoppelGANger** – Time-series GAN
- **SDV** – Unified SDK with multiple models

Comparison of Selected Synthetic Data Tools:

Tool	Data Type	Strengths	Challenges
CTGAN	Tabular	Handles skewed data	Hyperparameter sensitive
DoppelGANger	Time-series	Long sequences	High resource cost
SDV	Multiple	Easy to use	Generic results

Reference:

- ODSC, 2023: “9 Open-Source Tools to Generate Synthetic Data”

Problem Statement

Core Research Problem:

Many tools exist, but there is a lack of comprehensive, objective comparison—especially across data types and evaluation metrics.

- No universal best tool → performance depends on data type, use case, and context.
- Current comparisons are either anecdotal or tool-specific.

🧠 *"How can we evaluate and compare synthetic data generation tools fairly?"*

Research Task & Workflow

Research Task:
To perform a fair, structured comparison of synthetic data generation tools across data types and evaluation methods.

My approach includes following steps:

1. **Tool Selection**
   Identify open-source tools for tabular and time-series data (e.g., CTGAN, DoppelGANger, SDV).

2. **Dataset Preparation**
   Choose benchmark datasets suitable for testing each tool's capabilities.

3. **Data Generation**
   Use each tool to generate synthetic versions of the datasets.

4. **Evaluation**
   Assess generated data by:

   - Statistical similarity (e.g., distributions, correlations)
   - ML utility (e.g., performance on classification tasks)
   - Resource usage (e.g., runtime, memory)

5. **Comparison & Interpretation**
   Compare results side by side and interpret strengths/limitations of each tool.

6. **Recommendations**
   Provide practical guidelines on when and where to use each tool.

Aim & Objectives

Aim:
To conduct a comparative analysis of open-source synthetic data generation tools for tabular and time-series data.

Objectives:

- Analyze selected tools' designs and features.
- Evaluate output quality: statistical similarity, ML utility, performance.
- Benchmark on shared datasets.
- Make practical recommendations.

References

- Endres et.al., 2022 – Synthetic Data Generation: A Comparative Study
- Jordon et al., 2018 – Hide-and-Seek Privacy GAN
- Goncalves et al., 2020 – Synthetic Data Survey
- MITRE, 2021 – SDV documentation
- ODSC, 2023 – Medium article

Thank you for listening