
PHAROS
THE GREEK AI FACTORY
AI4Health Topic
Training Series
Course 8
Generative Modeling in Medical Data
MARCH 23, 2026 | 12:00 EET | ONLINE
PHAROS

Who we are

Intelligence Systems Lab

SCHOOL OF INFORMATICS

ARISTOTLE UNIVERSITY OF THESSALONIKI

- Ioannis Vlahavas, Professor

- Vasileios Kochliaridis, PhD candidate

- Zoi Katsantoni, MSc graduate

Healthcare paradox

- It is recognized that healthcare faces a great paradox, which at the same time is an attractive challenge:
  - On one hand, the development of effective AI for personalized medicine that requires vast amounts of diverse patient data
  - On the other hand, the legitimately binding strict privacy regulations such as GDPR (EU) and HIPAA (USA)

So, how can we innovate responsibly while protecting patients?

- Synthetic data is the answer
- They are transforming healthcare into data-driven ecosystem

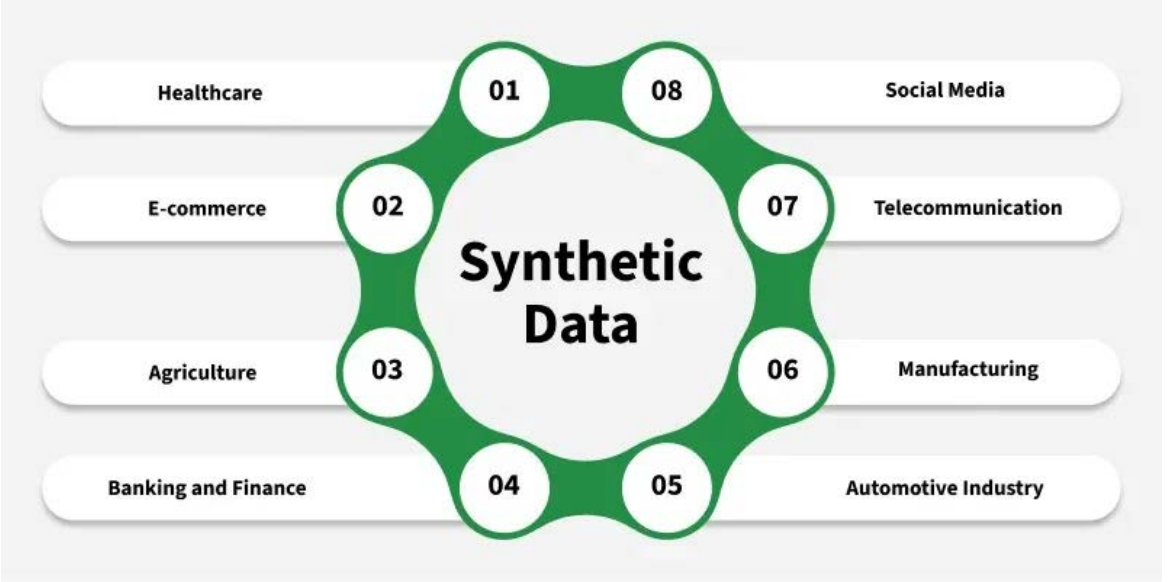
How Synthetic Data Can Transform Healthcare

- Synthetic data supports AI training, data sharing, and privacy protection
- However, there are some concerns about
  - Bias, transparency, accountability, as well as about
  - EU AI Act regulatory framework that leaves gaps regarding validation of synthetic data and continuously evolving AI systems

Definitions

- Synthetic data refers to artificially generated information that mirrors the statistical patterns of real patient data without containing identifiable information

  - They look and behave like the real data for research, but you cannot trace them back to any individual.
  - They are created using generative AI and allow researchers to train AI models without accessing restricted, sensitive patient records.



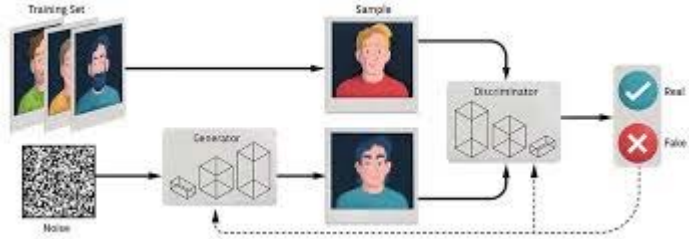
Benefits of Synthetic Data

- Enhanced Privacy: Enables data sharing and research without compromising sensitive patient information
- Improved Generalizability: Allows for the creation of larger, more diverse datasets that reduce bias, especially in rare diseases or specialized populations (e.g., pediatrics)
- Data Availability: It fills gaps when real data is unavailable, scarce, or difficult to obtain, fostering innovation in medical software development

Techniques for Synthetic Data

There are two main categories of techniques:

- Statistical Methods. Simple and explainable but only for tabular data
  - Random Over/Under Sampling, SMOTE, Borderline-SMOTE and ADASYN
- Machine Learning (Neural networks) - Generative Adversarial Networks (GANs) – Low Efficiency - 2D images
  - cGANs (Conditional GANs)– Guided generation
  - CTGAN (Conditional Tabular GANs)
  - DCGAN (Deep Convolutional GANs)
  - WGAN (Wasserstein GAN)
  - StyleGAN3 (NVIDIA)
  - R3GAN



- Diffusion and Guided Diffusion Models that overcome traditional GAN models most of the times

Evaluation metrics of synthetic data (1/3)

The evaluation of synthetic data measures the quality across three main pillars:

- Fidelity: Statistical resemblance to real data.
  Evaluates Correlation and Distribution similarity.
  Assesses if synthetic data behaves like real data, using both visual and numerical techniques
- Evaluated by 3 metrics
  - PSNR: Peak Signal-to-Noise Ratio. Is used to calculate reconstruction error and signal fidelity
  - SSIM: Structural Similarity Index metric. A methodology for estimating the perceived quality of digital images and videos
  - FID: Fréchet Inception Distance. The most widely used and reliable metric. Compares the degree of similarity between the distribution of the generated images and the actual ones
- The combined study of these three indicators allows us to draw safe conclusions about both the sharpness and the naturalness of the final images

Evaluation metrics of synthetic data (2/3)

- Utility:

  Measure the usefulness of synthetic data
- Clinical Evaluation by Experts (Visual Turing Test)
  - Quantitative metrics (FID, SSIM) are useful indicators, but they cannot replace clinical judgment
- Evaluation of Usefulness in Classification Tasks (Machine Learning Utility)
  - A critical step in certifying the value of synthetic data is its practical application in the training of diagnostic algorithms
  - Train on Synthetic, Test on Real (TSTR methodology)
    - We train a classifier exclusively with the synthetic data and evaluate it in real patient data

Evaluation metrics of synthetic data (3/3)

- Privacy:

  These ensure synthetic data does not memorize or leak original data

- Privacy Preserving Metrics
  - Given the sensitive nature of medical data, it's critical to quantify the risk of information leakage.
    - To apply metrics to test the model's memorization of training data, as well as to assess resilience to re-identification attacks.
  - The aim is to ensure that the generated data is adequately anonymized and not they are just copies or just a slightly altered version of real data of real patients.
  - Distance to Closest Record (DCR), Linkability, Nearest Neighbor Distance Ratio (NNDR)
    - For each data (record) s in the synthetic dataset (S), we calculate its distance to every data (record) in the original, real dataset (R).
    - The DCR for a synthetic data $s \in S$ is the minimum of these distances:
    - $\mathrm{DCR}(s) = \min_{r \in R}$ *distance(s,r)*

Conclusion

- The growing field of synthetic data holds promise for applications such as privacy protection and data augmentation, as well as its ability to accelerate development and democratize research
- Synthetic data complements but does not replace real data, but
  - Provide large, pre-labeled, and diverse datasets needed to train robust AI models, reducing overfitting and enhancing accuracy
  - By generating data rather than collecting it, organizations can fill gaps in underrepresented datasets, improve model performance, and ensure privacy compliance (e.g., GDPR, HIPAA)
  - Address Data Scarcity and Bias
  - Accelerate Clinical Research
  - They allow Cost-Effective Research-Innovation saving time and resources

Course Description

- This course consists of 2 parts
- The 1st Part, introduces key concepts and methodologies for working with healthcare and medical data and generating synthetic datasets for research and Machine Learning applications
- Finally, the 2nd Part, discusses applications, benefits, and challenges of synthetic data in healthcare.
- More specifically:

Part 1: Data generation

- It begins with an overview of healthcare data types and characteristics, including structured tabular data from electronic health records and medical imaging data such as MRIs and EEGs
- The course presents traditional statistical approaches for synthetic data generation and data balancing, including techniques such as SMOTE, ADASYN, and other resampling methods commonly used to address class imbalance
- It also covers modern Deep Learning approaches for Generative Modeling, including autoencoders and GANs for synthesizing realistic tabular and visual medical data
- Finally, it demonstrates applications, benefits, and challenges of synthetic data in healthcare

Part 2: Generative Models in MRI Data

- The secont Part, presents the development of Generative AI methodologies for the synthesis of real three-dimensional medical data, specifically volumetric magnetic resonance imaging (MRI) of the brain
  - The initial approach, which was based on a purely 3D model (3D R3GAN), failed to converge
- To overcome this obstacle, the research turned to the use of two-dimensional architectures, examining two different techniques for integrating the third dimension
  - Channel Stacking (when applied to 2D R3GAN, brought about the first major success)
  - Slice-based Generation using Positional Conditioning (Conditional R3GAN improved the quality
- The final and most successful architecture turned out to be Guided Diffusion with Positional Conditioning
- In summary, this presentation demonstrates that direct 3D generation of medical data remains a challenge under limited resources.
  - On the contrary, reducing the problem to a 2D level with Positional Conditioning is an effective strategy


PHAROS
THE GREEK AI FACTORY
AI4Health Topic
Training Series
Course 8
Generative Modeling in Medical Data
MARCH 23, 2026 | 12:00 EET | ONLINE
PHAROS