Generating Private and Fair Long-Sequenced Longitudinal Healthcare Records

Md Fahim Sikder, Resmi Ramachandranpillai and Fredrik Heintz

Reasoning and Learning Lab, Department of Computer and Information Science (IDA), Linköping University

Abstract

Generating long-sequenced longitudinal healthcare records is critical as it has numerous potential applications. Long-sequenced longitudinal data allow us to better understand and find patterns from the data. However, privacy concerns make it challenging to share the data, and real-world data is not bias-free. Generative Adversarial Networks (GAN) have been used to synthesize healthcare records, but the high dimensionality of these data makes them challenging to generate. From these motivations, we are working on a diffusion-based model that is capable of generating long-sequenced, fair, and private healthcare records.

Introduction

Overview

This project is part of an ongoing PhD study. In the PhD study, we aim to propose methods for generating high-quality and long-sequenced time series data and maintaining privacy and fairness in the model. Also, we aim to propose methods to evaluate the quality of synthetic long-sequence time-series data.

Motivation

- Long-sequenced time-series data gives more information than the short-sequenced data.
- Most time-series generative models are Generative Adversarial Networks (GAN) [1, 3] and training GAN is challenging as well as it prones to mode-collapse problem.
- Transformers architecture can capture long-term dependencies.

Overall Research Challenges

How to generate long-sequenced and high-dimensional time-series data?

How to make the data/model privacy-preserved?

How to make the data/model fair?

How to evaluate the data/model with respect to privacy, fairness and fidelity?

Method

Our proposed model works in the following steps:

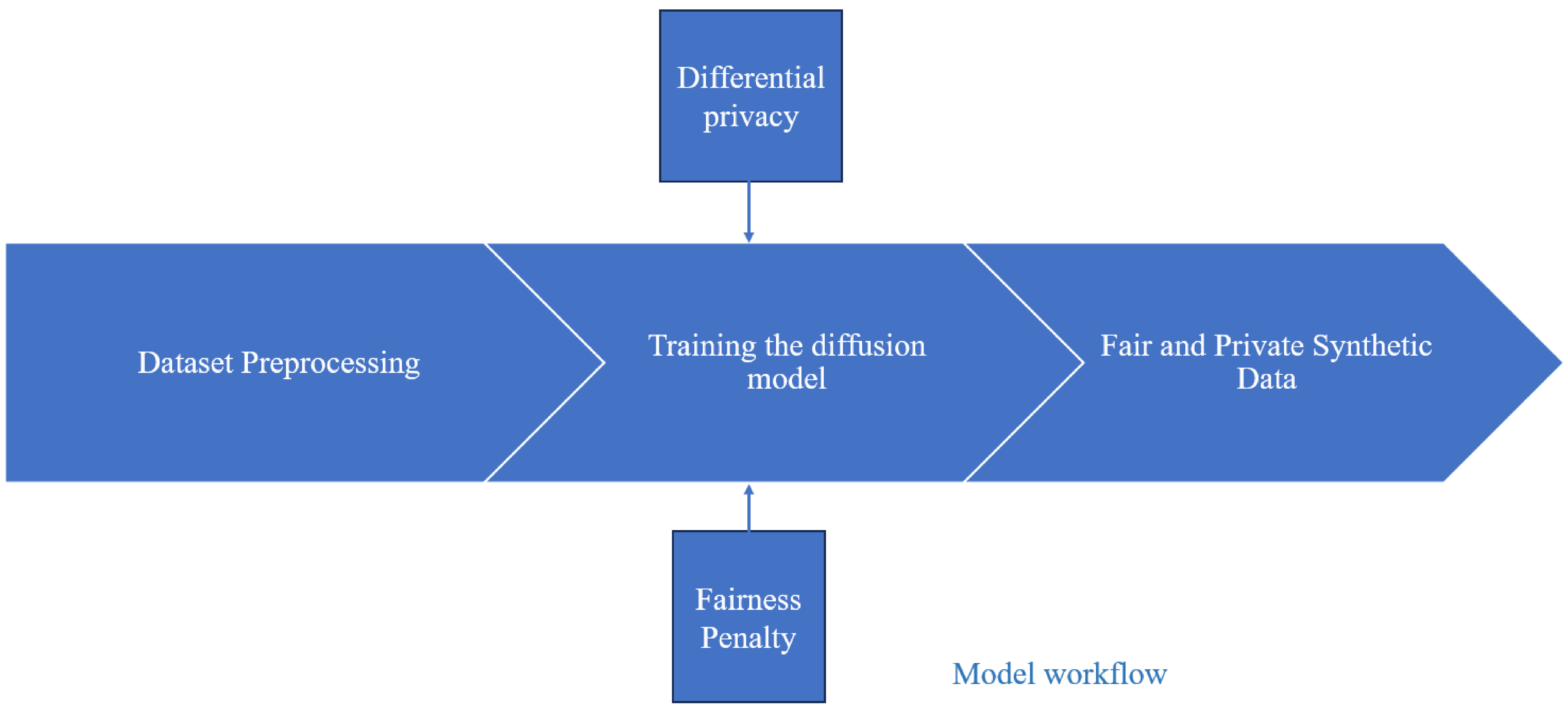
- **Diffusion Forward step**: We add noise to the data until the data become pure Gaussian noise.
- **Diffusion Backward step**: We use a Transformer-Encoder [4] based neural network to denoise the data and approximate the original data distribution.
- **Adding fairness penalty:** We add fairness penalty to the diffusion backward step to ensure fair data generation.
- **Train in DP manner:** We also train the whole process in differential privacy (DP) manner to ensure private data generation.

Why Private and Fair Synthetic Data

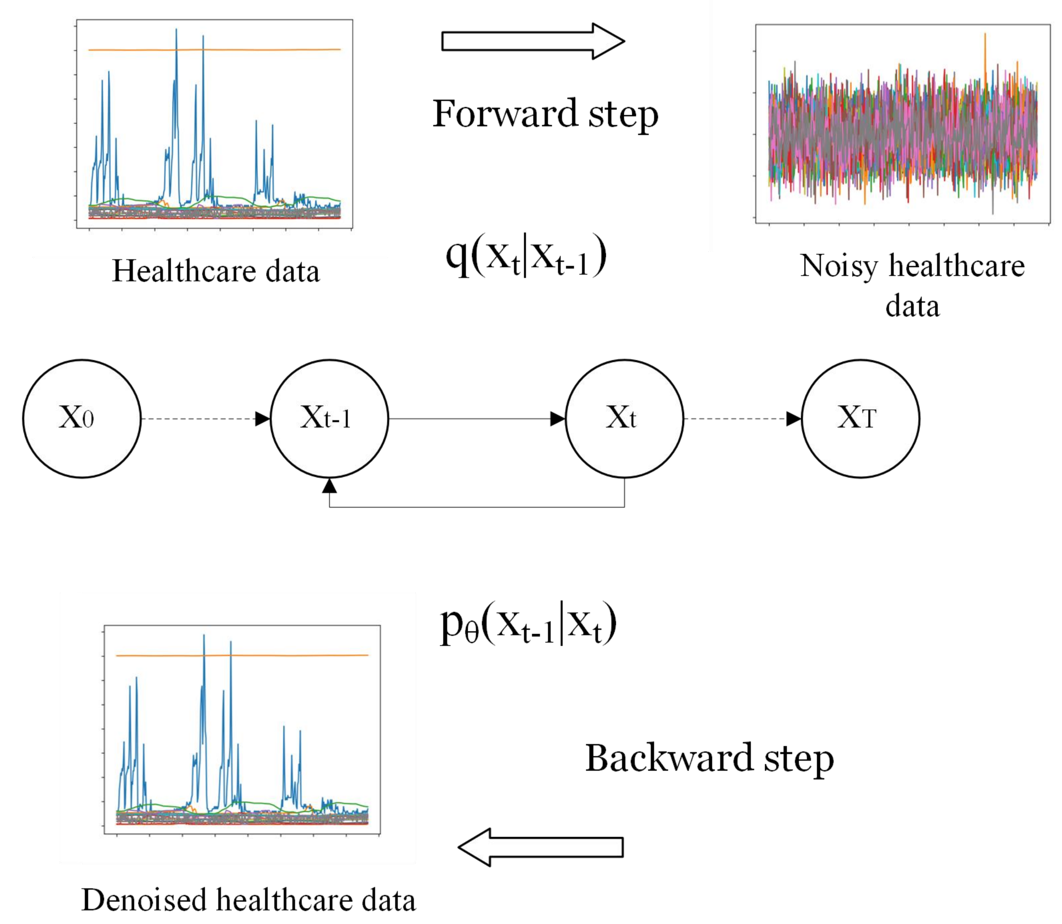
- Private synthetic healthcare data can improve the quality of reasearch and development without compromising patient's privacy.
- Fair synthetic healthcare data can mittigate the bias issue in the real data.

Our Contribution

- Our model is capable of generating long-sequenced longitudinal healtcare records.
- Generates fair and private data.
- We add fariness penalty to ensure fair data generation.
- Usage of Diffusion and Transformers overcome the mode-collapse problem of generative models.



Model workflow



Diffusion workflow

Evaluation

- We use two benhmarking healtcare datasets (MIMIC-III & MIMIC-IV).
- We use **Gender** as the sensitive attribute.
- Visual Evaluation
  - PCA & t-SNE Plots
- Empirical Evaluation
  - **Fidelity**: LDS, Jensen-Shannon Divergence (JSD), **α**-precision [5]
  - **Diversity**: **β**-recall [5]
  - **Check mode collapse**: Coverage [5]
  - **Predictive Analysis**: LPS, +5 Steps Ahead
- Fairness Evaluation
  - Demographic Parity
  - Predictive Parity
- Privacy Evaluation
  - AUROC
  - LDS
  - Authenticity [5]

Acknowledgment

This work was funded by the Knut and Alice Wallenberg Foundation, the ELLIIT Excellence Center at Linköping-Lund for Information Technology, and TAILOR - an EU project with the aim to provide the scientific foundations for Trustworthy AI in Europe. The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

References

[1] Xu, Tianlin, et al. "Cot-Gan: Generating Sequential Data via Causal Optimal Transport." In the Proceedings of *Neural Information Processing Systems (NeurIPS)*, 2020.

[2] Ho, Jonathan, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models."In the Proceedings of *Neural Information Processing Systems (NeurIPS)*, 2020.

[3] Yoon, Jinsung, Daniel Jarrett, and Mihaela Van der Schaar. "Time-Series Generative Adversarial Networks." In the Proceedings of *Neural Information Processing Systems (NeurIPS)*, 2019

[4] Vaswani, Ashish, et al. "Attention is All You Need." In the Proceedings of *Neural Information Processing Systems (NeurIPS)*, 2017

[5] Alaa, Ahmed, et al. "How Faithful is Your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models." In the Proceedings of *International Conference on Machine Learning (ICML)*, 2022.

Poster

LINKÖPING UNIVERSITY

Department of Computer and Information Science (IDA)
Artificial Intelligence and Integrated Computer Systems (AIICS)