Master Thesis Proposal Nr:

Topic: Synthetic data for clinical machine learning

Data is the foundation of clinical machine learning, yet real-world data often poses significant limitations. It may be privacy-sensitive, imbalanced, unrepresentative, unfair, or entirely unavailable [1]. To overcome these challenges, synthetic data is increasingly being used to fill gaps, particularly where annotated medical data is scarce [2].

Synthetic data can be generated using perturbations with accurate forward models (models that simulate outcomes for specific inputs), physical simulations, or AI-driven generative models [2]. However, this process introduces trade-offs: while synthetic data can increase dataset size, it may compromise data quality by introducing unrealistic elements [1].

For instance, in the context of renal cell carcinoma, the chromophobe subtype is rare, comprising only 5% of all cases [2]. By generating synthetic histology images of this subtype and using them as additional training data for a convolutional neural network, detection accuracy for this rare subtype can be improved [2].

The usefulness of synthetic data has been further validated in studies such as the one by Azizi et al. [3]. Their paper, *"Can synthetic data be a proxy for real clinical data? A validation study,"* demonstrated that synthetic clinical trial datasets could produce similar analytical results and conclusions as real datasets. Using a survival analysis of patients with colon cancer as an example, they showed that synthetic data can serve as an effective substitute for real data in certain applications [3].

Despite these successes, generating synthetic datasets is not without challenges. Implicit issues such as fairness and representativeness in data distribution can affect model performance, potentially leading to biases and discriminatory practices in real-world applications.

This thesis will explore the current landscape of synthetic data generation methods for clinical data analysis. It will include an extensive literature review, comparative testing of various synthetic data generation techniques, and an evaluation of their feasibility, advantages, and pitfalls. The primary focus will be on electronic health records (text and tabular data).

Additionally, the thesis will analyze the ethical considerations surrounding synthetic data, highlighting its benefits and addressing challenges such as data quality, fairness, and representativeness in medical research. By doing so, this work aims to provide a comprehensive understanding of the role of synthetic data in advancing clinical machine learning.

Tasks:

1. **Literature Review**
   - Research existing synthetic data generation methods in clinical data.
   - Select at least two clinical datasets to which you create synthetic data for.
2. **Method Testing**
   - Implement at least three synthetic data generation methods (e.g., synthcity [4])
3. **Method Evaluation**
   Possible metrics could be data usefulness, which evaluates the extent to which synthetic data resemble the statistical properties of the original data, and information disclosure, which measures how much of the real data can be shown by the synthetic data [3].

4. **Ethical Analysis**
   - Investigate the ethical considerations of using synthetic data in clinical research (e.g. [5,6]).
   - Highlight privacy benefits and potential biases introduced by synthetic data generation. Try to use concrete examples from your results
5. **Documentation and Thesis Writing**

The thesis must contain a detailed description of all developed and used algorithms as well as a profound result evaluation and discussion. The implemented code has to be documented and provided. Extended research on literature, existing patents and related work in the corresponding areas has to be performed.

Advisors: Dr. Emmanuelle Salin, Arijana Bohr M. Sc., Prof. Dr. Bjoern Eskofier
Student: Ludwig Roggenkamp
Start—End: 01.01.2025—01.06.2025

References

[1] Van Breugel, B. and van der Schaar, M., 2023. Beyond privacy: Navigating the opportunities and challenges of synthetic data. *arXiv preprint arXiv:2304.03722*.
[2] Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F. and Mahmood, F., 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, *5*(6), pp.493-497.
[3] Azizi, Z., Zheng, C., Mosquera, L., Pilote, L. and El Emam, K., 2021. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ open*, *11*(4), p.e043497.
[4] https://github.com/vanderschaarlab/synthcity
[5] Van Breugel, B., Kyono, T., Berrevoets, J. and Van der Schaar, M., 2021. Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, *34*, pp.22221-22233.
[6] Hao, S., Han, W., Jiang, T., Li, Y., Wu, H., Zhong, C., Zhou, Z. and Tang, H., 2024. Synthetic data in AI: Challenges, applications, and ethical implications. *arXiv preprint arXiv:2401.01629*.