GENERATING SYNTHETIC DATA FOR JUSTICE INFORMATION SYSTEMS

Eda Nur Deniz
HCAIM, HU

Rick Kosse
Justid supervisor

Academic advisor
HCAIM, HU

1 ABSTRACT

There is a significant demand for data sharing both within and outside the Department of Justice Information System (Justid). The privacy rules outlined in the General Data Protection Regulation (GDPR) heavily restrict the sharing of personal data. This restriction extends to sharing data with scientific institutions and other governmental bodies. Moreover, developers and researchers who require access to this data face limitations due to privacy regulations, resulting in complex workflows and increased susceptibility to errors. Synthetic data presents a viable solution to this problem. By not containing personal information, synthetic data bypasses the restrictions imposed by the GDPR. Consequently, the shareability of data is enhanced, enabling developers to create effective applications. Simultaneously, researchers can analyze the data without compromising individual privacy.

2 INTRODUCTION

Organizations handling sensitive data, such as Justid, face significant challenges in data accessibility due to strict GDPR regulations [1]. These restrictions limit the ability of developers and researchers to utilize real-world data for innovation and analysis. Traditional anonymization techniques often fail to ensure both privacy and data utility.

To address this problem, this research will explore (open-source) deep learning techniques for generating synthetic data [2] to evaluate their effectiveness in balancing privacy, security, and usability for legal and scientific applications.

Research question:
What (open source) methods are available for generating synthetic data, and to what extent are these methods capable of masking the original data?

Hypothesis
Synthetic datasets of tabular data, generated using advanced deep learning techniques and privacy-preserving measures, will be non-traceable to the original data while maintaining utility. These datasets will also comply with legal frameworks like GDPR more effectively than traditional anonymization methods.

3 LITERATURE REVIEW

Figure 1 and Table 1 illustrate the key trends and findings from previous research on techniques used for generating synthetic data. These visualizations highlight the evolution of methods and their effectiveness in addressing data accessibility and privacy challenges.

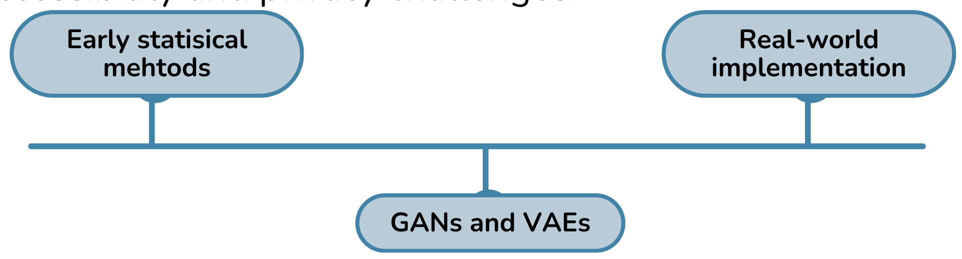

Figure 1: The evolution of synthetic data approaches [3-5]

Method	Privacy level	Data accuracy	Use case
Bootstrapping	High	Low	Basic models
Bayesian Networks	Moderate	Moderate	Risk analysis
GANs (TGAN, CTGAN)	Low	High	Generating complex data

Table 1: Comparing different methods that were the key in generating synthetic data [3-6]

4 RESEARCH METHODOLOGY

This research follows an experimental and exploratory approach to generate synthetic data, ensuring both statistical validity and privacy compliance. Figure 2 illustrates the key steps required to complete the proces.

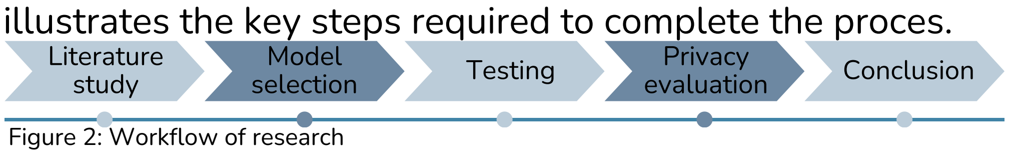

Figure 2: Workflow of research

This research bridges the gap between data accessibility and privacy protection. By addressing technical challenges and human-centered and ethical aspects, it allows for responsible data sharing within legal constraints.

5 PRELIMINARY CONSIDERATIONS

This research improves generating synthetic data by balancing privacy and accuracy. The findings may enhance machine learning models and support GDPR-compliant data sharing for the Justice Information Services.

6 CONCLUSION

This research is crucial in addressing privacy concerns by exploring the generation of synthetic data, facilitating secure data sharing and application development while ensuring GDPR compliance. This approach has the potential to drive innovation and collaboration within Justid, advancing data science while safeguarding privacy.

References.
[1] He, S., & Gao, T. (2022). High-Resolution Mapping of Global Surface Water Extent Using Sentinel-1 Synthetic Aperture Radar Data. arXiv, 2205.03257. https://arxiv.org/pdf/2205.03257
[2] European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L119, 1-88. https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng
[3] Wajid Shafiq. (2024). Optimizing Organizational Performance: A Data-Driven Approach in Management Science. Bulletin of Management Review, 1(2), 31–40. Retrieved from https://bulletinofmanagement.com/index.php/Journal/article/view/48
[4] Bourou, S., El Saer, A., Velivassaki, T.-H., Voulkidis, A., & Zahariadis, T. (2021). A Review of Tabular Data Synthesis Using GANs on an IDS Dataset. *Information*, *12*(9), 375. https://doi.org/10.3390/info12090375
[5] Brenninkmeijer, B. (2019). On the Generation and Evaluation of Tabular Data using GANs. ResearchGate. https://www.researchgate.net/publication/344227988_On_the_Generation_and_Evaluation_of_Tabular_Data_using_GANs
[6] Dwork, C., McSherry, F., Nissim, K., Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In: Halevi, S., Rabin, T. (eds) Theory of Cryptography. TCC 2006. Lecture Notes in Computer Science, vol 3876. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11681878_14

Acknowledgements. The HCAIM (the Human-Centred AI Master's Programme) Project is Co-Financed by the Connecting Europe Facility of the European Union Under Grant №CEF-TC-2020-1 Digital Skills 2020-EU-IA-0068. This poster was created as part of the Blended Intensive Programme organised under the Erasmus + Programme of the European Union.