Master's Thesis Seminar Paper

Enhancing Privacy in Machine Learning through Teacher-Guided Synthetic Data Generation: A Modified PATE Framework

Student: Timo Maksan
Advisor: Ao.univ.Prof. Dr. Andreas Rauber

June 14, 2025

1 Abstract

The increasing use of machine learning in sensitive domains such as healthcare and finance raises pressing concerns about data privacy. Traditional privacy-preserving approaches like the PATE (Private Aggregation of Teacher Ensembles) framework still rely on auxiliary datasets, creating potential vectors for indirect leakage. To address this, the thesis proposes a novel modification to the PATE framework that replaces label aggregation with synthetic data generation: each teacher model, trained on a private data subset, produces synthetic data that collectively trains a student model without revealing real inputs. This enhances privacy by eliminating the need for real or auxiliary data in student training. The methodological approach follows the CRISP-DM process model and integrates principles from empirical AI research to ensure rigorous experimentation and reproducibility. Evaluation will focus on privacy guarantees, model utility, and distributional fidelity, enabling a comprehensive assessment of the proposed framework's effectiveness.

2 Selected Research Method

This seminar paper investigates the methodological foundation applied in the thesis project, which proposes a novel variation of the PATE (Private Aggregation of Teacher Ensembles) framework using teacher-guided synthetic data generation. The selected methodology is grounded in the **CRISP-DM process model**, complemented by principles from empirical AI research, systematic experimentation, and reproducibility guidelines.

CRISP-DM as the Core Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) framework was selected as the central process model for the thesis due to its wide adoption, domain independence, and structured guidance across all phases of the machine learning lifecycle [9, 8]. The six iterative stages—Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment—provide a comprehensive blueprint for implementing, documenting, and critically evaluating the proposed approach.

Schröer et al. [9] confirm the continued relevance of CRISP-DM in data-driven research and offer best practices for each phase, including structured templates for presenting methodological decisions. Rauber [8] further reinforces CRISP-DM's utility in ensuring reproducibility and process transparency in AI-based research projects.

Empirical and Quasi-Experimental Design

The research follows a quasi-experimental setup to assess the impact of the synthetic data aggregation approach compared to standard PATE methods. This includes a controlled comparison of model performance, privacy budget, data efficiency, and distributional fidelity.

Cohen [4] outlines how empirical methods can be effectively used to test hypotheses in AI by emphasizing structured experimental pipelines, documentation, and repeatability. This aligns with the experiment design approach recommended by Blockeel and Vanschoren [2], who propose maintaining traceable experiment databases and treating models and results as scientific artifacts. These ideas are echoed by Chugh [3], who stresses the importance of planning, versioning, and reporting in experimental data science.

Theory-Guided and Domain-Aware Structuring

The thesis also integrates ideas from theory-guided data science [5], which emphasize the incorporation of domain knowledge into data mining workflows. This is particularly relevant for ensuring that synthetic data generated by teacher models preserves relevant feature distributions and class boundaries. Angelov and Gu [1] complement this view by promoting hybrid modeling strategies that combine machine learning with knowledge-based assumptions, especially useful for guiding experimental expectations.

Relevance to PATE and Differential Privacy Evaluation

The PATE framework by Papernot et al. [7] serves as the technical foundation for the thesis. Their methodology of teacher ensemble learning and privacy budgeting using the moments accountant is reused and extended in this work. The experimental setup retains the basic structure of their work while replacing label aggregation with synthetic data generation—thus introducing a new methodological dimension within the established paradigm.

Finally, the research structure also draws from Rauber's seminar slides [8], which encourage process-model-driven planning (e.g., CRISP-DM) and empirical rigor to ensure the traceability and reproducibility of AI research.

3 Literature Summary

This section presents the key methodological insights from selected literature that inform the research design of the thesis. The focus lies on how each work contributes to the understanding, structuring, and implementation of empirical machine learning projects, particularly within the CRISP-DM framework and experimental AI.

Papernot et al. (2017)

Papernot et al. [7] introduced the PATE framework, a foundational approach in privacy-preserving machine learning. Their work provides a reference methodology for training ensembles of teacher models on disjoint data subsets and using differentially private aggregation to train a student model. The thesis adapts this methodology by replacing label aggregation with synthetic data generation, making Papernot et al. a baseline for both structure and evaluation metrics, including differential privacy accounting.

Angelov and Gu (2019)

Angelov and Gu [1] emphasize the need for empirical modeling strategies that balance learning from data with knowledge-based insights. Their staged approach to machine learning experimentation—ranging from initial heuristics to hypothesis validation—is particularly relevant for evaluating the proposed synthetic data method and understanding data efficiency.

Blockeel and Vanschoren (2007)

Blockeel and Vanschoren [2] propose the use of experiment databases to improve transparency and reproducibility in machine learning research. Their framework promotes tracking all model configurations, evaluation results, and experimental decisions, which is directly applied in this thesis through structured logging and experiment versioning.

Cohen (1995)

Cohen [4] provides foundational insights into empirical AI, including how to design quasi-experiments, control confounding factors, and document results rigorously. The thesis borrows from his recommendations on defining measurable research questions, forming hypotheses, and aligning evaluation strategies to empirical goals.

Karpatne et al. (2017)

Karpatne et al. [5] advocate for theory-guided data science, where domain knowledge is integrated with machine learning workflows. In the thesis, this is applied by aligning synthetic data generation with known distributional characteristics (e.g., class balance, feature ranges), enhancing interpretability and trust in model behavior.

Schröer et al. (2021)

Schröer et al. [9] conducted a systematic literature review of CRISP-DM applications, extracting best practices for each of the six process phases. Their structured template for applying and documenting CRISP-DM is directly applied in the thesis, particularly for reporting business understanding, modeling approaches, and evaluation metrics.

Chugh (2022)

Chugh [3] provides practical guidelines for designing experiments in data science, stressing the importance of setting clear objectives, versioning experiments, and using automation to track configurations. These practices are adopted in the thesis to ensure reproducibility and minimize bias across runs.

Rauber (2024)

Rauber [8] emphasizes the integration of process models such as CRISP-DM into AI research projects to enhance reproducibility, documentation quality, and methodological rigor. His lecture materials also highlight the importance of model traceability, audit trails, and aligning evaluation strategies with initial business goals—all of which are adopted in the thesis methodology.

4 Application to Thesis Work

The methodological insights drawn from the literature are operationalized in the thesis through a structured application of the CRISP-DM process model and principles of empirical, reproducible research. This section describes how each CRISP-DM phase is adapted to the proposed privacy-preserving synthetic data framework.

Business Understanding

Following the recommendations of CRISP-DM [9, 8], the project begins by identifying the key problem: reducing the privacy leakage risk of the PATE framework by eliminating its reliance on auxiliary or real student training data. The goal is to enable private model training by generating synthetic data from each teacher. This aligns with the research questions defined in the thesis proposal [6] regarding privacy, model utility, and data efficiency.

Data Understanding

The primary data used in this thesis includes benchmark datasets such as MNIST and Fashion-MNIST. Data characteristics such as class balance, distribution skew, and feature ranges are analyzed statistically and visually, following the best practices outlined by Schröer et al. [9] and Chugh [3]. This phase also includes inspecting synthetic outputs from teachers to ensure they reflect realistic and diverse data properties, as recommended in theory-guided modeling literature [5].

Data Preparation

Data is partitioned into disjoint subsets for each teacher model. This aligns with the original PATE method [7], where teachers do not share data. Cleaning and preprocessing steps include normalization, format conversion, and potential augmentation. Synthetic data is generated per teacher using conditional generative models or noise-perturbed sampling. The preparation logic is documented with configuration scripts, in line with Blockeel and Vanschoren's experimental traceability guidelines [2].

Modeling

Each teacher is trained on its respective private data subset using a consistent model architecture (e.g., CNNs). Synthetic data samples are then generated and aggregated into a dataset for student training. Multiple configurations (e.g., sample size, noise levels, generation method) are evaluated to observe effects on model accuracy and privacy budget, as per the empirical design practices outlined in Cohen [4] and Angelov and Gu [1]. All experiments are versioned and logged.

Evaluation

Student models trained on synthetic data are evaluated for accuracy, precision, recall, and other utility metrics. Additionally, privacy budgets are computed using differential privacy accounting, such as the Moments Accountant [7, 4]. Distributional metrics (e.g., KL divergence, feature histograms) are used to assess fidelity between real and synthetic data distributions.

To contextualize performance, the student model trained on synthetic data is benchmarked against models trained using traditional PATE aggregation, as well as recent state-of-the-art privacy-preserving methods from the literature. This comparative evaluation ensures that the proposed method is not only effective in isolation but also competitive in terms of privacy-utility trade-offs.

Deployment

While actual deployment is beyond the thesis scope, the CRISP-DM deployment phase is addressed by preparing reproducible code repositories, experiment configuration files, and documentation to enable reproducibility and future use. Schröer et al. [9] note that deployment is often underrepresented in research, and this work addresses that gap by offering a fully documented pipeline for further experimentation and real-world adaptation.

5 Conclusion

This seminar paper has outlined the methodological foundation for a master thesis that proposes a privacy-preserving modification to the PATE framework by introducing teacher-guided synthetic data generation. The core methodology follows the CRISP-DM process model, which has proven to be an effective and adaptable framework for structuring machine learning projects [9, 8].

The thesis also integrates principles from empirical AI research, quasi-experimental design, and theory-guided modeling to ensure that the proposed approach is both scientifically rigorous and practically reproducible. By drawing on a diverse set of methodological sources [4, 2, 5, 1], the work is grounded in best practices for data preparation, model evaluation, and documentation.

Furthermore, the experimental design incorporates strategies for reproducibility and transparency, such as version-controlled configurations and detailed evaluation metrics [3]. This methodological rigor not only strengthens the scientific contribution of the thesis but also addresses common shortcomings in machine learning research—particularly with respect to deployment, traceability, and evaluation.

Overall, the selected methodology supports the thesis objectives by enabling a clear, repeatable, and transparent research process. It ensures that the privacy guarantees, utility, and data efficiency of the synthetic data approach can be critically assessed and potentially generalized to other privacy-sensitive domains.

References

[1] Plamen P. Angelov and Xiaowei Gu. *Empirical Approach to Machine Learning.* Springer, 2019.

[2] Hendrik Blockeel and Joaquin Vanschoren. Experiment databases: Towards an improved experimental methodology in machine learning. *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 6–17, 2007.

[3] Vidhi Chugh. The importance of experiment design in data science. `https://www.kdnuggets.com/2022/08/importance-experiment-design-data-science.html`, 2022. Accessed: 2025-06-15.

[4] Paul R. Cohen. *Empirical Methods for Artificial Intelligence.* MIT Press, 1995.

[5] Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop R. Ganguly, and Vipin Kumar. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10):2318–2331, 2017.

[6] Timo Maksan. Enhancing privacy in machine learning through teacher-guided synthetic data generation: A modified pate framework. In *Master's Thesis Proposal*, Vienna, Austria, 2025.

[7] Nicolas Papernot, Martin Abadi, Úlfar Erlingsson, Ian Goodfellow, and Kunal Talwar. Semi-supervised knowledge transfer for deep learning from private training data. *International Conference on Learning Representations (ICLR)*, 2017.

[8] Andreas Rauber. Research methods in data analysis: Process models and reproducibility. In *Lecture Slides, TU Wien*, 2024.

[9] Christoph Schröer, Felix Kruse, and Jorge Marx Gómez. A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534, 2021.