Master's Thesis Proposal

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

May 29, 2025

1 Motivation & Problem Statement

The increasing use of machine learning in sensitive domains such as healthcare, finance, and education has heightened concerns about data privacy. Conventional approaches often rely on centralized datasets, exposing users to potential privacy breaches and regulatory non-compliance. To address this, privacy-preserving machine learning (PPML) techniques have emerged, with frameworks like *Federated Learning* and *Private Aggregation of Teacher Ensembles (PATE)* gaining notable attention for their ability to learn from decentralized data without direct access to raw inputs.

PATE, in particular, offers a promising balance between privacy and utility by training multiple teacher models in disjoint subsets of sensitive data and aggregating their (noisy) predictions to label a student model [11]. However, this method still requires a real or auxiliary dataset to be labeled and used for student training—which introduces a potential vector for indirect data leakage, especially if the auxiliary dataset overlaps or correlates with sensitive distributions.

To push privacy boundaries further, this thesis proposes a modification to the PATE framework: rather than using aggregated labels, each teacher will generate synthetic data that mirrors the statistical properties of its private subset. The union of these datasets will be used to train a student model, eliminating the need to expose real data or labels altogether. This approach introduces new challenges, particularly regarding data utility, representativeness of synthetic samples, and quantification of privacy guarantees under differential privacy (DP) frameworks [8, 1].

Furthermore, it remains unclear whether synthetic data generation can preserve the diversity and decision boundaries of the original data sufficiently to train competitive models. While techniques for synthetic data generation and differential privacy have advanced, their application in ensemble-based privacy-preserving training remains underexplored.

Therefore, this work is motivated by the need to:

- Minimize the exposure of real or auxiliary data during student model training;
- Assess the data efficiency and privacy trade-offs of synthetic-only learning;

- Examine the theoretical and practical implications of using synthetic teacher outputs in a distributed privacy-preserving learning setup.

This novel direction addresses a critical research gap at the intersection of privacy, synthetic data, and ensemble learning—with the potential to set new standards for safe machine learning in privacy-sensitive contexts.

2 Aim of the Work

The primary aim of this thesis is to design, implement, and evaluate a novel variation of the PATE (Private Aggregation of Teacher Ensembles) framework that replaces the traditional label aggregation step with a synthetic data generation approach. In this modified setup, each teacher model generates synthetic data based on its private subset, and these synthetic datasets are then combined to train a student model. This aims to improve the privacy guarantees of the PATE framework by eliminating the need for real or auxiliary data during student training.

To assess the effectiveness and implications of this approach, the thesis will investigate the following research questions:

- **RQ1:** To what extent does training a student model exclusively on synthetic data generated by teacher models preserve predictive performance compared to the standard PATE framework?
- **RQ2:** How much privacy budget (in terms of differential privacy parameters ϵ, δ) is consumed by the proposed approach, and how does it compare to standard PATE?
- **RQ3:** How much synthetic data is required to reach a defined performance threshold on a downstream classification task?
- **RQ4:** To what degree does the combined synthetic dataset replicate the distributional properties (e.g., feature coverage, label balance) of the original sensitive data?

The thesis will also briefly explore optional extensions such as a feedback loop in which teacher models adjust synthetic outputs based on student misclassification patterns, potentially enabling adaptive refinement of the synthetic dataset.

By answering these questions, the work seeks to contribute to the field of privacy-preserving machine learning by proposing a framework that strengthens data protection without severely compromising model performance — thus offering a practical alternative for scenarios where data sensitivity is paramount.

3 Methodological Approach

This thesis follows a structured, empirical approach grounded in the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology [13], tailored to suit machine learning research as described in empirical AI literature [7], experiment methodology frameworks [3], and theory-guided data science principles [9]. The goal is to ensure reproducibility, rigorous evaluation, and scientific robustness while investigating a novel synthetic-data-based extension to the PATE framework.

3.1 Research Design

The study employs a quasi-experimental design to evaluate how replacing the standard PATE aggregation mechanism with teacher-generated synthetic data impacts privacy and model performance. Key comparisons are drawn between:

- The standard PATE method with noisy label aggregation;
- The proposed method with synthetic data aggregation.

Experiments are conducted under controlled conditions to isolate the effect of the synthetic data generation step. Each research question is addressed through measurable evaluation criteria.

3.2 CRISP-DM Framework

The CRISP-DM process guides the methodology across the following stages [13]:

1. **Business Understanding:** Clarify the goal of enhancing privacy in distributed ML while preserving performance through synthetic data aggregation.
2. **Data Understanding:** Analyze the disjoint private datasets used to train teacher models and ensure representativeness and diversity.
3. **Data Preparation:** Normalize, encode, and clean the private subsets to support fair synthetic generation. Synthetic data will be validated for quality and representational fidelity.
4. **Modeling:**
   - Train teacher models on disjoint subsets.
   - Each teacher generates a synthetic dataset using a generative model (e.g., GANs or VAEs).
   - Combine these to form a student training set.
   - Train the student model without access to real or auxiliary data.
5. **Evaluation:** Measure model performance, privacy guarantees, data efficiency, and representational quality (see below).

6. **Documentation and Reproducibility:** Maintain experiment databases and traceable logs for all model configurations, datasets, and code pipelines [3, 5].

3.3 Evaluation Strategy

Each research question is addressed through targeted evaluation metrics:

- **RQ1 (Performance Comparison):** Evaluate the student model trained on synthetic data vs. standard PATE using accuracy, precision, recall, F1-score, and ROC-AUC.
- **RQ2 (Privacy Budget):** Use differential privacy accounting tools (e.g., moments accountant [11] or Rényi DP) to measure (ϵ, δ) and compare privacy budget consumption.
- **RQ3 (Data Efficiency):** Vary the size of the synthetic training data (e.g., 10%, 25%, 50%, 100%) and observe the resulting performance. Plot learning curves to determine the minimal quantity needed to reach a predefined threshold (e.g., 90% of baseline accuracy) [2].
- **RQ4 (Distributional Fidelity):** Assess how well the combined synthetic datasets replicate the distribution of the original data using:
  - Statistical measures (e.g., Jensen-Shannon divergence, Earth Mover's Distance),
  - Class distribution and feature correlation comparison,
  - Visualizations via dimensionality reduction (e.g., PCA or t-SNE plots) [9].

3.4 Exploratory Component (Optional)

An exploratory sub-task may evaluate a feedback loop: teacher models detect misclassification patterns in the student and generate additional targeted synthetic examples to refine the dataset iteratively. This will be piloted only if time permits and evaluated qualitatively.

4 Structure of the Work

1. **Introduction**

   - Motivation and problem statement
   - Research objectives and scope
   - Research questions
   - Overview of the proposed approach
   - Thesis outline

2. **Related Work**

    - Federated Learning and decentralized training
    - Differential Privacy and privacy accounting methods
    - The Private Aggregation of Teacher Ensembles (PATE) framework
    - Synthetic data generation techniques (GANs, VAEs, diffusion models)
    - Hybrid and alternative privacy-preserving ML frameworks

3. **Methodological Approach**

    - Empirical research design
    - Application of the CRISP-DM process model
    - Reproducibility considerations and experiment documentation
    - Evaluation dimensions and corresponding metrics
    - Alignment of methodology with research questions

4. **Design and Implementation**

    - Training teacher models on disjoint private subsets
    - Synthetic data generation (model selection, configuration, constraints)
    - Aggregation and quality control of synthetic datasets
    - Student model training pipeline
    - Integration of differential privacy mechanisms

5. **Evaluation**

    - Experimental setup and datasets
    - Model performance comparison (standard PATE vs. proposed method)
    - Privacy budget consumption analysis
    - Data efficiency through synthetic sample scaling
    - Distributional similarity and feature-level fidelity analysis

6. **Discussion**

    - Summary of findings per research question
    - Trade-offs between privacy, utility, and efficiency
    - Observed limitations and sources of variance

- Implications for real-world deployment
- Outlook on adaptive refinement mechanisms (feedback loops)

7. **Conclusion and Future Work**

- Summary of contributions
- Reflections on the methodology and findings
- Recommendations for future research
- Suggestions for real-world adoption in regulated industries

5 State-of-the-Art

The Private Aggregation of Teacher Ensembles (PATE) framework has been a foundational approach in privacy-preserving machine learning, offering strong differential privacy guarantees by aggregating outputs from multiple teacher models trained on disjoint subsets of sensitive data. Since its inception, several extensions and adaptations have been proposed to enhance its applicability, utility, and privacy guarantees.

5.1 PATE-GAN

PATE-GAN integrates the PATE framework with Generative Adversarial Networks (GANs) to generate synthetic data that maintains differential privacy. In this approach, the discriminator is trained using differentially private mechanisms, ensuring that the generator produces synthetic data without compromising individual privacy. Experiments have demonstrated that PATE-GAN can produce higher quality synthetic data compared to other differentially private GANs, such as DPGAN [8].

5.2 G-PATE

G-PATE extends the PATE framework by employing an ensemble of teacher discriminators to train a student generator in a differentially private manner. This method leverages private gradient aggregation techniques to ensure privacy during the training process, enabling the generation of high-dimensional data with improved utility under strict privacy budgets [10].

5.3 Individualized PATE

Traditional PATE applies a uniform privacy budget across all data points, which may not reflect the varying privacy preferences of individuals. Individualized PATE addresses this by allowing different privacy budgets for different data points, using techniques like upsampling and weighting. This approach enhances the utility of the resulting models while respecting individual privacy requirements [4].

5.4 Hot PATE

Hot PATE is designed for tasks with inherently diverse outputs, such as generative language models. It modifies the aggregation mechanism to handle distributions over outputs rather than single labels, preserving the diversity of responses while maintaining differential privacy guarantees. This extension is particularly beneficial for in-context learning scenarios [6].

5.5 PATE-CTGAN

PATE-CTGAN combines the PATE framework with Conditional Tabular GAN (CTGAN) to generate differentially private synthetic tabular data. This integration is particularly suited for structured data, offering a balance between data utility and privacy. By applying PATE to CTGAN, the approach mitigates issues like mode collapse and enhances the quality of synthetic data for downstream tasks [14].

5.6 Handling Heterogeneous Data

Standard PATE assumes homogeneously distributed data across teacher models. However, in real-world scenarios, data distributions can be heterogeneous. Recent work has proposed modifications to the teacher training process, such as incorporating teacher averaging and update correction, to address high variance in teacher updates due to data heterogeneity. These adjustments lead to improved consensus among teachers and enhanced student model performance [12].

6 Relevance to the Curricula of Business Informatics

This thesis contributes to the field of Business Informatics by addressing a central challenge at the intersection of data privacy, machine learning, and information systems design. The work draws upon core competencies of the curriculum, including data-driven decision-making, system architecture, and privacy-aware analytics.

By enhancing the PATE framework with a synthetic data generation mechanism, the thesis tackles both technical and organizational issues relevant to secure data processing in business environments. It reflects the increasing demand for privacy-preserving machine learning in domains such as healthcare, finance, and public administration—sectors where compliance with data protection regulations (e.g., GDPR) is critical.

References

[1] Nesime Tatbul Abay, Yichen Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy-preserving synthetic data release using deep learning. *arXiv preprint arXiv:1801.01594*, 2018.

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

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

[4] Franziska Boenisch, Christopher Mühl, Roy Rinberg, Jannis Ihrig, and Adam Dziedzic. Individualized pate: Differentially private machine learning with individual privacy guarantees. *arXiv preprint arXiv:2202.10517*, 2022.

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

[6] Edith Cohen, Benjamin Cohen-Wang, Xin Lyu, Jelani Nelson, Tamas Sarlos, and Uri Stemmer. Hot pate: Private aggregation of distributions for diverse tasks. *arXiv preprint arXiv:2312.02132*, 2023.

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

[8] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. *arXiv preprint arXiv:1806.03384*, 2018.

[9] 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.

[10] Yunhui Long, Boxin Wang, Zhuolin Yang, Bhavya Kailkhura, Aston Zhang, Carl A. Gunter, and Bo Li. G-pate: Scalable differentially private data generator via private aggregation of teacher discriminators. *arXiv preprint arXiv:1906.09338*, 2019.

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

[12] Nicolas Papernot, Shuang Song, Ilya Mironov, Aditi Raghunathan, Kunal Talwar, and Úlfar Erlingsson. Scalable private learning with pate. In *International Conference on Learning Representations*, 2018.

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

[14] Microsoft SmartNoise. Create privacy-preserving synthetic data for machine learning with smartnoise, 2021.