Logan Choi
Thesis Plan

Proposal

The healthcare industry increasingly relies on data and machine learning to enhance diagnostic accuracy and patient outcomes. However, organizations must ensure patient data privacy and comply with HIPAA (Health Insurance Portability and Accountability Act) regulations, even while leveraging machine learning capabilities. Homomorphic encryption (HE) offers a promising solution by allowing computations to be performed directly on encrypted data, ensuring that sensitive information remains secure throughout the analysis process.

This study explores the feasibility of using TenSEAL, a Python-based library, to integrate HE with machine learning techniques such as logistic regression. The goal is to determine whether encrypted medical data can be securely analyzed without compromising privacy or accuracy. Specifically, this study has the following research questions:

1. **How does homomorphic encryption impact the predictive accuracy of machine learning models on healthcare datasets?**
2. **What are the differences in performance between the Brakerski/Fan-Vercauteren (BFV) and Cheon-Kim-Kim-Song (CKKS) encryption schemes in terms of model accuracy?**
3. **What are the computational costs (encryption time, RAM usage, memory overhead) associated with using HE in machine learning?**

The study will answer it through the following objectives:

1. **Compare Model Performance**: Evaluate predictive accuracy differences between encrypted and plaintext data to assess HE's impact on machine learning models.
2. **Analyze Encryption Schemes**: Compare the Brakerski/Fan-Vercauteren (BFV) and Cheon-Kim-Kim-Song (CKKS) encryption schemes across multiple healthcare datasets, focusing on their impact on the accuracy of the machine learning model as the primary evaluation metric.
3. **Measure Computational Overhead**: Collect and analyze encryption time, RAM usage, and memory overhead for each encryption scheme.

By conducting these evaluations, this study aims to provide insights into the practicality of homomorphic encryption for secure healthcare applications.

Schedule

The thesis will be structured in two phases across Spring and Fall 2025. The Spring quarter will focus primarily on experimentation and data collection, while the Fall quarter will be dedicated to analyzing findings and writing the final thesis paper. I will also be enrolled in the thesis writing course during Fall 2025 to support the writing and presentation process.

Spring 2025 (Experimentation Phase):

- **Week 1-2:** Begin by finalizing and submitting the detailed thesis plan. During this period, I will run preliminary experiments using mock datasets to evaluate the computational overhead of different homomorphic encryption schemes (BFV and CKKS). The goal is to support the third research question by measuring encryption time, RAM usage, and memory footprint to assess scalability.
- **Week 3-5:** Use the same logistic regression models developed earlier to run experiments on three biological datasets sourced from Kaggle. These experiments will focus on comparing predictive performance across plaintext, BFV-encrypted, and CKKS-encrypted versions of the datasets, addressing both the first and second research questions.
- **Week 6-9:** Begin developing and documenting a custom dataset using the MIMIC IV clinical database. Once the dataset is ready, run encryption and machine learning experiments on it. This phase will demonstrate a complete process from raw data extraction and preparation to encryption and predictive modeling. Including this example will help strengthen the paper by showing a real-world application of the full workflow.
- **Week 10:** Wrap up the quarter by organizing all experimental results, observations, and notes. The focus will be on setting up a clear structure for analysis and writing in the upcoming quarter, ensuring everything is in place for the final stretch and eventual defense.

Fall 2025 (Research Paper Phase):

- **Week 1-2:** Organize all relevant data, research papers, and scientific articles related to the thesis. Use this time to build a strong storyboard and outline the structure of the thesis paper.
- **Week 3-4:** Begin writing the introduction, literature review, and abstract. Meet with committee members to discuss the narrative flow, key talking points, and overall direction of the paper.
- **Week 5-6:** Create charts and graphs from the data collected during the experimentation phase. Focus on writing the methodology, results, and conclusion sections. Share progress with the committee and gather feedback for revisions.
- **Week 7-8:** Start preparing for the final defense. Work on the presentation and begin practicing the delivery to ensure clarity and confidence.
- **Week 9-10:** Complete the slides and finalize preparations for the defense. Present to the committee, submit the final version of the thesis paper, and complete the requirements for graduation.

(Note: The schedule is flexible if there are any unforeseen circumstances.)

Quality Criteria & Metrics

Throughout the entire thesis, a range of quality checks and evaluation metrics will be employed to ensure the research is thorough, reproducible, and impactful. These metrics are categorized into three main areas: model performance, encryption efficiency, and overall research quality.

Model Evaluation Metrics:

To evaluate model effectiveness, accuracy, precision, recall, and F1-score will be recorded for both encrypted and plaintext datasets. These metrics will help ensure that the use of homomorphic encryption does not significantly degrade model performance. Additionally, a full step-by-step process will be documented to serve as a reference example for applying encryption in a machine learning pipeline.

Encryption Metrics:

To assess the practical impact of encryption, measurements will be taken for encryption time, computational overhead, memory usage, and scalability. Encryption time refers to the time required to encrypt datasets before training, while computational overhead focuses on how much longer it takes to train and evaluate models on encrypted data compared to plaintext. Memory usage will be tracked to understand the differences in RAM and storage requirements, and scalability will be tested by analyzing how encryption performance changes as the dataset size increases.

Research Quality:

To ensure that the research is credible and methodologically sound, several steps will be taken. All code and processes will be clearly documented to support reproducibility, allowing others to replicate the results. The experiments will also be designed for robustness, ensuring consistent outcomes across different runs and scenarios. Finally, fair comparisons will be maintained by using the same models and hyperparameters for both encrypted and unencrypted datasets, helping to avoid bias and ensuring that observed differences can be attributed solely to the encryption layer.

Software Processes

To ensure the success of this thesis, a structured approach to software development and research organization is essential. Version control will be maintained through Git and GitHub, allowing for efficient code management, collaboration, and tracking of changes over time. All code will follow best practices, including thorough documentation, consistent formatting, and clear inline comments that explain the logic and workflow of each script. Experiment tracking will be a priority, with results carefully logged and organized to maintain clarity and enable easy comparison between runs.

Special attention will be given to data handling, particularly when working with sensitive information from the MIMIC-IV database. Proper data governance protocols will be followed to prevent any leakage of personal or identifying information. Finally, reproducibility will be a core focus of the research. All steps—from data preprocessing to model training and evaluation—will be clearly documented to ensure that other researchers can replicate the experiments and validate the results with confidence.