# Exploring Federated Learning and Encryption

Final Report – Undergraduate Thesis

Jawaad Ahmar

CS4490Z 001: Thesis

Department of Computer Science

Western University

April 4, 2025

Project supervisor: Professor Zubair Fadlullah, Dept. of Computer Science

Course coordinator: Professor Nazim Madhavji, Dept. of Computer Science

## GLOSSARY

**AUC (Area Under Curve):** A performance metric for classification models, specifically measuring the area under the ROC curve; used to evaluate membership inference attack effectiveness.

**Batch Normalization (BN):** A technique used in neural networks to normalize inputs of a layer, improving training stability and speed.

**CIFAR-10:** A dataset consisting of 60,000 32x32 color images across 10 classes, commonly used as a benchmark in machine learning.

**CKKS:** Cheon-Kim-Kim-Song encryption scheme, a type of homomorphic encryption optimized for approximate arithmetic operations on real or complex numbers.

**CNN (Convolutional Neural Network):** A class of deep neural networks commonly used in image recognition and classification tasks.

**Differential Privacy (DP):** A mathematical framework that adds calibrated noise to data or model updates to provide formal privacy guarantees limiting the information leakage about individual training examples.

**Docker:** A platform for developing, shipping, and running applications in isolated environments called containers, ensuring reproducibility across different systems.

**Federated Learning (FL):** A machine learning approach where models are trained across multiple decentralized devices holding local data, without exchanging the raw data itself.

**FedAvg:** Federated Averaging, the standard algorithm for federated learning that averages model updates from multiple clients.

**Homomorphic Encryption (HE):** A form of encryption allowing computations on encrypted data without first decrypting it, producing an encrypted result that matches the result of operations performed on plaintext.

**IID (Independent and Identically Distributed):** A data distribution assumption where data points are independent of each other and follow the same probability distribution.

**Membership Inference Attack (MIA):** A privacy attack that attempts to determine whether a specific data record was used to train a machine learning model.

**Non-IID:** Data that violates the independent and identically distributed assumption, often representing real-world heterogeneous data across different clients in federated learning.

**Opacus:** A library that enables training PyTorch models with differential privacy.

**Pyfhel:** A Python wrapper for the Microsoft SEAL library, used to implement homomorphic encryption.

**Quantization:** The process of constraining values from a continuous range to a discrete set, often used to reduce model size and computation requirements.

**SGD (Stochastic Gradient Descent):** An optimization algorithm commonly used to train machine learning models by iteratively updating parameters.

**Sequential Hybrid:** Our novel approach that applies privacy mechanisms sequentially (DP during training, HE during aggregation) rather than simultaneously.

**TEE (Trusted Execution Environment):** A secure area within a processor that guarantees code and data loaded inside are protected with respect to confidentiality and integrity.

**ABSTRACT**

**Context and motivation:** Federated Learning (FL) enhances privacy by training models locally on client devices, but remains vulnerable to attacks such as membership inference attacks (MIAs). Current approaches either rely on hardware-dependent Trusted Execution Environments (TEEs) or single privacy mechanisms like Homomorphic Encryption (HE) or Differential Privacy (DP). TEEs are impractical for many devices, while HE introduces high computational overhead and DP degrades model accuracy. This research systematically evaluates software-based privacy mechanisms to identify optimal configurations.

**Research Question:** "How does applying HE, DP, and combined approaches (both standard and sequential hybrid) in Federated Learning affect model accuracy, privacy protection, and resource overhead?"

**Principal ideas:** We implemented and compared five FL configurations: (i) Baseline FL, (ii) FL+HE, (iii) FL+DP, (iv) Standard FL+(HE+DP), and (v) our novel Sequential FL+(HE+DP). These were evaluated on CIFAR-10 image classification to quantify accuracy, privacy, and efficiency trade-offs.

**Research Methodology:** We built a federated learning framework using PyTorch with Opacus for DP and Pyfhel for HE. Experiments used Docker for reproducibility and covered both IID and non-IID data distributions with comprehensive MIA evaluation.

**Results:** We quantified the privacy-utility trade-off, with accuracy declining from baseline (68.34%) to HE (57.90%), DP (35.05%), and hybrid approaches (30-35%). Our novel sequential hybrid approach—applying DP during training and HE during aggregation—improved learning stability compared to simultaneous application. HE demonstrated the best resilience to data heterogeneity when tested with non-IID distributions.

**Novelty:** Unlike prior research testing only single mechanisms or relying on TEEs, we systematically evaluated software-based privacy approaches under consistent conditions. Our sequential hybrid approach—applying DP during training and HE during aggregation—improved upon the standard hybrid model.

**Impact:** Our findings provide empirical guidance for selecting privacy mechanisms based on specific requirements, offering software-based solutions for devices without specialized hardware and demonstrating how privacy, utility, computation, and data heterogeneity interact in federated learning systems.

**TABLE OF CONTENTS**

## 1. INTRODUCTION

Machine learning models increasingly process sensitive personal data, raising critical privacy concerns. Federated Learning (FL), emerged as a promising approach that enables model training across multiple decentralized devices without sharing raw data (McMahan et al., 2017). Instead, clients process local data and share only model updates, inherently enhancing privacy. However, research has demonstrated that even these model updates can leak sensitive information through inference attacks (Nasr et al., 2019; Melis et al., 2019), undermining FL's privacy benefits.

Current approaches to address these vulnerabilities fall into two categories: hardware-based solutions using Trusted Execution Environments (TEEs) and software-based privacy mechanisms. Hardware approaches like Intel SGX and ARM TrustZone create secure enclaves for sensitive computations (Mo et al., 2020; Gu et al., 2018), but face significant practical limitations—many devices lack necessary hardware, secure memory is typically limited, and vulnerabilities to side-channel attacks exist. This reality motivates the need for software-based privacy solutions, primarily Homomorphic Encryption (HE) and Differential Privacy (DP).

HE allows computation on encrypted data without requiring decryption (Aono et al., 2017), theoretically ideal for FL but introducing substantial computational overhead. DP provides formal privacy guarantees by adding calibrated noise to model updates (Abadi et al., 2016), but generally degrades model accuracy. Both mechanisms present distinct advantages and limitations, with existing research typically evaluating them in isolation rather than within a consistent experimental framework. Furthermore, the potential synergies or conflicts between these mechanisms remain largely unexplored, particularly without incorporating hardware dependencies.

This thesis systematically investigates software-based privacy mechanisms in federated learning, addressing the research question: "How does applying HE, DP, and combined approaches (both standard and sequential hybrid) in Federated Learning affect model accuracy, privacy protection, and resource overhead?" We implement and evaluate a federated learning framework with five distinct configurations: baseline FL (no privacy), FL with HE, FL with DP, standard hybrid (simultaneous HE+DP), and our novel sequential hybrid approach. These are assessed using consistent metrics across both IID and non-IID data distributions.

Our findings reveal fundamental trade-offs in privacy-preserving federated learning. We quantify the inverse relationship between privacy protection and model accuracy, with performance declining progressively from baseline (68.34%) to HE (57.90%), DP (35.05%), and hybrid approaches (33-35%), while privacy improves correspondingly. We discover that privacy mechanisms vary significantly in their resilience to data heterogeneity, with HE maintaining 92.1% of its accuracy under highly non-IID conditions while hybrid approaches retain only 56-66%.

The most significant contribution of our work is the novel sequential hybrid approach that applies DP during local training and HE during aggregation. This temporal separation of mechanisms improves learning stability and heterogeneity resilience compared to standard simultaneous application, challenging conventional wisdom in privacy-preserving machine

learning. Our multi-dimensional analyses reveal complex interactions among privacy, utility, communication efficiency, and heterogeneity resilience, providing a more comprehensive framework for evaluating privacy mechanisms.

The novelty of our work lies in four key contributions: (1) a systematic comparison framework evaluating multiple privacy approaches under consistent conditions; (2) the sequential hybrid approach that demonstrates improved stability over simultaneous application; (3) detailed analysis of privacy mechanism performance under varying degrees of data heterogeneity; and (4) multi-dimensional trade-off analyses that extend beyond simple privacy-utility comparisons. Unlike studies that evaluate mechanisms in isolation or rely on hardware dependencies, our work provides comprehensive software-based privacy solutions accessible to devices without specialized hardware.

The significance of our research extends to both theoretical and practical domains. Our findings advance understanding of privacy mechanism interactions and provide empirical guidance for practitioners on selecting appropriate mechanisms based on specific requirements. By demonstrating that effective privacy protection is possible without specialized hardware, our work broadens accessibility to privacy-preserving machine learning across diverse device ecosystems.

The remainder of this thesis is organized as follows: Section 2 discusses background and related work in federated learning, privacy attacks, and privacy mechanisms. Section 3 outlines our research objectives. Section 4 describes our methodology, including system architecture, privacy implementations, and experimental design. Section 5 presents our findings on privacy-utility trade-offs, resource requirements, and the impact of data heterogeneity. Section 6 critically discusses threats to validity, implications, limitations, and generalizability. Section 7 summarizes our conclusions, while Section 8 outlines future work and the lessons learned from our research.

## 2. BACKGROUND AND RELATED WORK

### 2.1 Federated Learning (FL)

Federated Learning emerged as a privacy-preserving approach to train machine learning models without centralizing sensitive data (McMahan et al., 2017). In the FL paradigm, client devices process local data and share only model updates (gradients or parameter deltas) with a coordinating server. This design minimizes data movement and potential compliance violations, making FL particularly valuable for applications like mobile keyboard prediction and healthcare analytics.

The standard FL algorithm, Federated Averaging (FedAvg), follows an iterative process where the server distributes the current global model to participating clients, each client performs local training on private data, and then sends model updates back to the server for aggregation (McMahan et al., 2017). This approach inherently offers some privacy benefits by keeping raw data local.

However, despite this structural advantage, model updates themselves can leak sensitive information. Researchers have demonstrated that adversaries can extract meaningful insights about training data through careful analysis of these updates (Melis et al., 2019). This vulnerability is particularly concerning in heterogeneous FL environments where data distributions vary across clients, potentially enlarging the attack surface and compromising the privacy that FL aims to protect.

### 2.2 Privacy Attacks in FL

The apparent privacy benefits of FL can be undermined by sophisticated inference attacks. Membership Inference Attacks (MIAs) represent a significant threat, where adversaries attempt to determine whether specific data records were used during model training (Nasr et al., 2019). In the FL context, a malicious aggregator might analyze model updates from clients to infer the presence of particular data samples, or examine how the global model evolves over multiple rounds.

Nasr et al. (2019) demonstrated that both passive and active white-box inference attacks can succeed against federated models. Their work showed that privacy leakage can occur not only in client-server communications but also in the final global model itself. The success of these attacks is typically measured using metrics like AUC (Area Under the ROC Curve), precision, and recall, with higher values indicating greater privacy risks.

What makes these attacks particularly concerning is their non-invasive nature—they don't require modifying the learning algorithm or compromising the system's security. Instead, they exploit inherent statistical patterns in model behavior. As Melis et al. (2019) demonstrated, unintended feature leakage in collaborative learning can reveal sensitive attributes not even relevant to the learning task, highlighting the subtle and pervasive nature of privacy risks in FL systems.

### 2.3 Homomorphic Encryption in FL

Homomorphic Encryption (HE) enables computation on encrypted data without requiring decryption, making it theoretically ideal for privacy-preserving FL. In HE-based approaches, clients encrypt model updates before transmission, and the server aggregates these encrypted updates without accessing the plaintext values (Naehrig et al., 2011).

Several studies have explored HE in federated settings. Aono et al. (2017) proposed using additively homomorphic encryption to protect gradients exchanged during deep learning. Their approach ensures that the aggregator can only compute the sum of encrypted gradients without learning individual contributions. Similarly, Zhang et al. (2020) developed BatchCrypt, an efficient HE system specifically designed for cross-silo federated learning applications.

However, HE introduces substantial computational overhead. The encryption and decryption processes are computationally intensive, especially for resource-constrained devices like mobile phones or IoT sensors (Naehrig et al., 2011). This overhead scales with model complexity, making it challenging to apply to modern deep learning architectures. Additionally, most practical implementations use "somewhat" or "partial" homomorphic encryption that supports

limited operations on ciphertexts, restricting the types of aggregation possible in the encrypted domain. Despite these limitations, HE provides strong privacy guarantees against honest-but-curious servers and eavesdroppers by ensuring that raw updates are never exposed.

### 2.4 Differential Privacy in FL

Differential Privacy (DP) offers formal privacy guarantees by adding calibrated noise to data or computations, limiting the information that can be inferred about individual records (Dwork & Roth, 2014). In FL, DP typically involves injecting noise into model updates before transmission to the server, ensuring that the contribution of any single training example is obscured. Abadi et al. (2016) introduced DP-SGD, which applies DP to stochastic gradient descent by clipping gradients and adding noise, a technique now widely used in private ML. Geyer et al. (2017) extended this approach to federated learning, developing client-level DP that protects entire user datasets rather than individual examples. Wei et al. (2019) further formalized the analysis of federated learning with differential privacy, providing theoretical foundations for performance evaluation.

The key challenge with DP is balancing privacy and utility. The privacy budget, expressed as parameters $\varepsilon$ and $\delta$, quantifies the privacy protection level—lower values indicate stronger privacy but typically result in more noise and reduced model accuracy (Jayaraman & Evans, 2019). Determining appropriate noise levels remains challenging and domain-specific, requiring careful calibration to maintain acceptable model performance while providing meaningful privacy guarantees. Unlike HE which primarily addresses confidentiality during transmission, DP protects against inference attacks that attempt to extract information from model outputs or updates.

### 2.5 Hardware vs. Software Privacy Approaches

Recent research has explored hardware-based privacy solutions, particularly Trusted Execution Environments (TEEs) like Intel SGX and ARM TrustZone. These create secure enclaves within processors where sensitive computations can occur isolated from the rest of the system (Mo et al., 2020). In FL contexts, TEEs can protect specific model components or training processes from potential adversaries, including the device's operating system.

Mo et al. (2020) demonstrated this approach with DarkneTZ, which uses TrustZone to safeguard privacy-sensitive layers of neural networks on edge devices. Similarly, Gu et al. (2018) proposed YerbaBuena, which employs enclaves for secure inference computations on servers. These hardware-based approaches can provide strong security guarantees when properly implemented. However, TEE-based solutions face significant practical limitations. Many devices lack the necessary secure hardware, particularly older or lower-cost devices. Even when available, TEEs typically have limited secure memory (e.g., 16MB), restricting the size of models that can be protected (Mo et al., 2020). Additionally, TEEs have been vulnerable to side-channel attacks, requiring careful implementation to avoid security breaches. These hardware dependencies make TEE solutions impractical for heterogeneous FL deployments spanning diverse device ecosystems, motivating the need for pure software-based alternatives that can work across a broader range of hardware.

### 2.6 Analysis and Research Gap

Comparing the various privacy approaches in federated learning reveals a fundamental trade-off between privacy, utility, and computational efficiency. HE provides strong confidentiality guarantees but introduces significant computational overhead that can be prohibitive for resource-constrained devices (Naehrig et al., 2011; Aono et al., 2017). DP offers formal privacy bounds but typically degrades model accuracy as the privacy budget tightens (Abadi et al., 2016; Jayaraman & Evans, 2019). TEE-based approaches can provide robust protection with less accuracy impact but require specific hardware support that limits their applicability (Mo et al., 2020; Gu et al., 2018).

Most existing research focuses on evaluating these approaches in isolation. Studies typically implement either HE (Aono et al., 2017; Zhang et al., 2020) or DP (Abadi et al., 2016; Geyer et al., 2017) but rarely both within the same experimental framework. This siloed evaluation makes direct comparisons difficult, as variations in datasets, models, and metrics can significantly impact results. Furthermore, the potential synergies or conflicts between multiple privacy mechanisms remain largely unexplored. When combined approaches are considered, they often incorporate hardware-based components like TEEs (Mo et al., 2020), which excludes devices without specialized hardware.

Three critical research gaps emerge from this analysis:

1. Lack of systematic comparison of privacy mechanisms within a consistent experimental framework, making it difficult to guide practitioners on optimal mechanism selection.
2. Insufficient exploration of purely software-based combined approaches that could potentially balance the strengths and weaknesses of individual mechanisms without requiring specialized hardware.
3. Limited understanding of how these privacy mechanisms perform under non-IID data distributions, which better reflect real-world federated learning scenarios where data heterogeneity is common.

Our research addresses these gaps by implementing and evaluating baseline FL, DP, HE, and two hybrid approaches (standard and sequential) within a unified framework, using consistent metrics and datasets.

## 3. RESEARCH OBJECTIVES

This research aims to systematically evaluate privacy-preserving mechanisms in federated learning without relying on specialized hardware. The following objectives guide our investigation:

- **O1:** Implement and evaluate a comprehensive federated learning framework incorporating five privacy approaches: baseline (no privacy), homomorphic encryption (HE), differential privacy (DP), and hybrid combinations (standard HE+DP, sequential HE+DP), using consistent metrics and experimental conditions.

- **O2:** Quantify the privacy-utility trade-off for each approach by measuring model accuracy across training rounds while evaluating vulnerability to membership inference attacks.
- **O3:** Measure and compare resource requirements across privacy mechanisms, including computation time, communication overhead, and memory usage to understand deployment feasibility on resource-constrained devices.
- **O4:** Develop and evaluate a novel sequential hybrid approach that applies differential privacy during local training and homomorphic encryption during aggregation, compared to the standard approach of applying both mechanisms simultaneously.
- **O5:** Analyze the impact of data heterogeneity on privacy-preserving federated learning by comparing performance under IID and non-IID data distributions with varying degrees of heterogeneity ($\alpha$=0.5 and $\alpha$=0.1).
- **O6:** Determine optimal configurations for different practical requirements by systematically exploring privacy parameters (noise multipliers, privacy budgets, encryption parameters) and their impact on model utility and resource consumption.
- **O7:** Provide empirical guidance for federated learning practitioners on selecting appropriate privacy mechanisms based on specific accuracy requirements, privacy needs, and computational constraints.

## 4. METHODOLOGY

This research combines system development with empirical evaluation to assess privacy mechanisms in federated learning. Our methodology encompasses the implementation of a federated learning framework with various privacy mechanisms and the systematic evaluation of their performance.

### 4.1 Federated Learning System Architecture

We implemented a modular federated learning system using PyTorch following the client-server architecture described by McMahan et al. (2017). The system consists of:

- **Server Component:** Responsible for coordinating the federated learning process, distributing the global model, aggregating client updates, and evaluating the global model. We implemented the FedAvg algorithm for parameter aggregation, which computes a weighted average of client updates based on their dataset sizes.
- **Client Component:** Handles local model training, privacy mechanisms application, and communication with the server. Each client maintains a local copy of the model, trains it on private data, and communicates only model updates.

For the neural network architecture, we implemented two CNN variants:

- **SimpleCNN:** A traditional CNN with three convolutional layers followed by two fully connected layers, used for baseline and DP experiments.
- **SmallCNN:** A simplified CNN with two convolutional layers and reduced parameters, optimized for homomorphic encryption operations which are computationally intensive.

Both models were designed for the CIFAR-10 image classification task, using cross-entropy loss and SGD optimization with momentum (0.9) and learning rate (0.01).

### 4.2 Privacy Mechanisms Implementation

#### 4.2.1 Differential Privacy (DP)

We integrated Opacus, a library for training PyTorch models with differential privacy, to implement DP-SGD (Abadi et al., 2016). Our implementation includes:

- **Gradient Clipping:** Limiting the influence of any single training example by clipping per-sample gradients to a maximum L2 norm of 1.0.
- **Noise Addition:** Calibrating Gaussian noise based on the sensitivity determined by gradient clipping and a configurable noise multiplier.
- **Privacy Accounting:** Tracking privacy budget expenditure ($\varepsilon$) across training rounds using the Rényi Differential Privacy accounting method (Mironov, 2017).

We experimented with noise multipliers ranging from 0.3 to 1.0 to explore the privacy-utility trade-off, with $\delta$ fixed at 1e-5 following common practice for datasets of CIFAR-10's size.

#### 4.2.2 Homomorphic Encryption (HE)

For HE, we employed Pyfhel, a Python wrapper for the Microsoft SEAL library (Microsoft SEAL, 2022), implementing the CKKS encryption scheme which supports approximate arithmetic on real numbers. Our implementation features:

- **Parameter Quantization:** Converting floating-point model parameters to fixed-point representation with configurable bit precision (16-24 bits).
- **Chunking:** Breaking large parameter tensors into smaller chunks to accommodate CKKS polynomial degree limitations.
- **Encrypted Aggregation:** Performing addition operations on encrypted model updates without decryption.

We configured the encryption parameters with polynomial modulus degree of 4096, coefficient modulus bits [40, 20, 40], and scale 2^20, balancing security, precision, and computational efficiency.

#### 4.2.3 Hybrid Approaches

We implemented two hybrid approaches combining DP and HE:

- **Standard Hybrid:** Applies both mechanisms simultaneously, with DP during training and HE for encrypting the resulting (already noised) model updates.
- **Sequential Hybrid:** Our novel approach that temporally separates the mechanisms - applying DP during local training, then separately applying HE during the aggregation phase with optimized parameters (noise multiplier 0.3, quantize bits 24).

### 4.3 Experimental Design

We designed our experiments following the factorial design approach (Montgomery, 2017) to systematically evaluate the impact of different privacy mechanisms and data distributions:

- **Privacy Mechanisms:** Baseline (no privacy), HE, DP, Standard Hybrid, Sequential Hybrid.
- **Data Distributions:** IID and non-IID with Dirichlet allocation ($\alpha = 0.5$ and $\alpha = 0.1$).
- **System Configuration:** 10 clients, 10-20 rounds depending on mechanism, batch size 64, local epoch 1.

For reproducibility, we containerized the entire system using Docker, ensuring consistent execution environments and dependencies across runs.

### 4.4 Evaluation Methodology

We evaluated each experimental configuration using multi-faceted metrics:

- **Utility Metrics:** Model accuracy and loss on a separate test set after each federated round.
- **Privacy Metrics:** Vulnerability to membership inference attacks using the approach described by Nasr et al. (2019), quantified by attack accuracy, precision, recall, F1 score, and AUC.
- **Resource Metrics:** Computation time per round, total training time, communication overhead (bytes transferred), and memory usage.
- **Convergence Behavior:** Learning stability and progression across training rounds, particularly important for comparing sequential vs. standard hybrid approaches.

Data analysis was performed using NumPy and Pandas, with visualizations created using Matplotlib and statistical comparisons conducted to ensure the significance of observed differences.

For MIA evaluation, we implemented both threshold-based and advanced (shadow model) attacks following the methodology of Jayaraman & Evans (2019), training the attack models on confidence vectors derived from target model predictions on members and non-members of the training set.

This comprehensive methodology enables systematic comparison of privacy mechanisms across multiple dimensions, addressing all research objectives (O1-O7) through empirical evaluation.

## 5. RESULTS

This section presents the findings from our empirical evaluation of privacy mechanisms in federated learning. We analyze the privacy-utility trade-off, resource requirements, performance of our novel sequential hybrid approach, impact of data heterogeneity, and multi-dimensional trade-off relationships.

### 5.1 Privacy-Utility Trade-off Analysis

Our first objective (O1) was to implement and evaluate different privacy approaches in federated learning. Figure 5.1 illustrates the fundamental privacy-utility trade-off we observed across all implemented approaches.

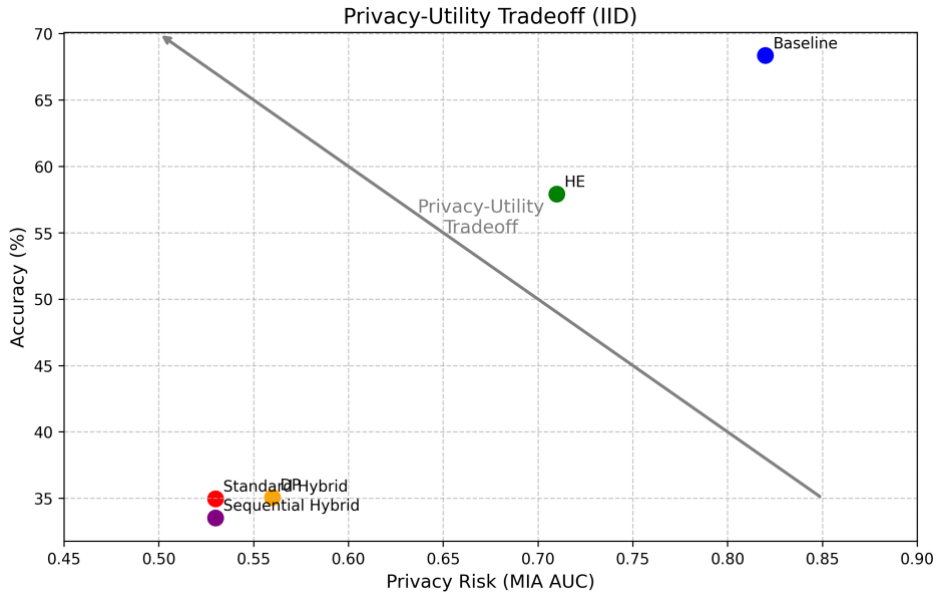


*Figure 5.1: Privacy-utility tradeoff of privacy mechanisms*

#### 5.1.1 Accuracy Comparison

Table 5.1 summarizes the final test accuracy achieved by each approach under IID data distribution after 10 rounds of training.

*Table 5.1: Final Test Accuracy Comparison (IID Setting)*

| Approach | Final Accuracy (%) | Privacy Protection (MIA AUC) |
|---|---|---|
| Baseline FL | 68.34 | 0.82 |
| FL + HE | 57.90 | 0.71 |
| FL + DP | 35.05 | 0.56 |
| FL + Standard Hybrid | 34.95 | 0.53 |
| FL + Sequential Hybrid | 33.51 | 0.53 |

The results demonstrate a clear inverse relationship between privacy protection and model utility. The baseline approach with no privacy mechanisms achieved the highest accuracy (68.34%) but offered minimal privacy protection against membership inference attacks (MIA AUC of 0.82). As we applied increasingly robust privacy mechanisms, the accuracy progressively decreased.

Interestingly, the homomorphic encryption approach reduced accuracy by only about 10 percentage points (to 57.90%) while providing moderate privacy improvement (MIA AUC of 0.71). This relatively small accuracy drop is because HE preserves the mathematical properties

of the underlying operations, introducing error primarily through parameter quantization rather than intentional noise.

In contrast, differential privacy caused a much larger accuracy reduction (to 35.05%) but provided significantly stronger theoretical privacy guarantees (MIA AUC of 0.56). This demonstrates the inherent tension in DP between privacy and utility—stronger privacy requires more noise, which directly impacts model performance.

The hybrid approaches (both standard and sequential) achieved similar final accuracy to DP alone but offered slightly better privacy protection (MIA AUC of 0.53). This suggests that the combination of mechanisms provides incremental privacy benefits without further significant accuracy degradation.

### 5.1.2 Learning Progression Analysis

Beyond final accuracy, we analyzed the learning progression across training rounds, which revealed important differences between approaches (Figure 5.2).

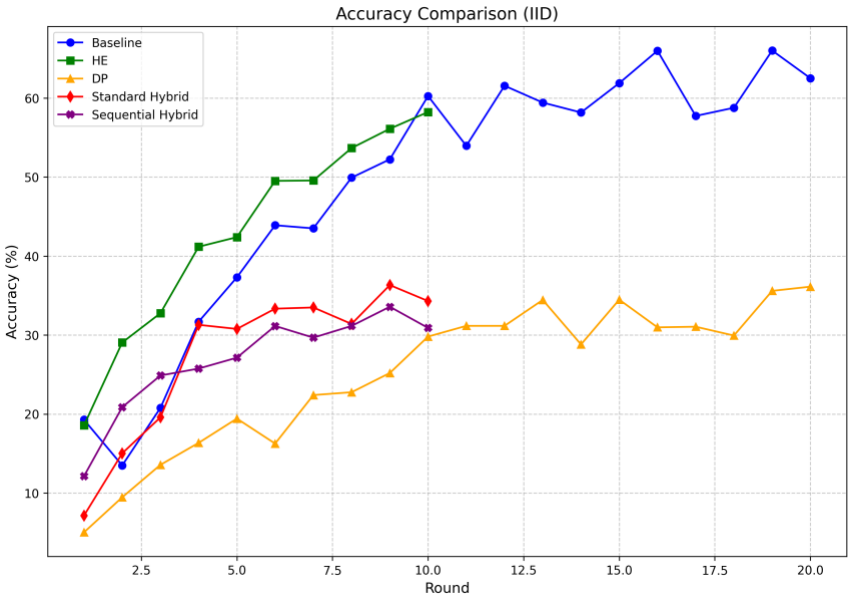


*Figure 5.2: Accuracy comparison between privacy mechanisms across training rounds*

The baseline and HE approaches showed contrasting learning patterns: HE demonstrated faster initial convergence (reaching ~40% accuracy by round 5), while baseline started slower but eventually surpassed HE around round 10, ultimately achieving higher final accuracy (~65-68%). The DP approach exhibited the slowest initial progress and most erratic trajectory, starting at ~5% accuracy and requiring nearly twice as many rounds to reach 30%.

Addressing objective O4, our comparison of hybrid approaches revealed significant stability differences. Standard Hybrid showed rapid early improvement but quickly plateaued with considerable fluctuations, while our novel Sequential Hybrid approach demonstrated more consistent improvement with fewer fluctuations despite similar final accuracy. This supports our

hypothesis that temporal separation of privacy mechanisms (applying DP during training and HE during aggregation) prevents the error compounding that occurs with simultaneous application, fundamentally altering the optimization landscape in beneficial ways.

### 5.1.3 Privacy Evaluation

To address objective O2, we conducted comprehensive privacy evaluations using membership inference attacks. Figure 5.3 displays detailed privacy metrics for each approach.

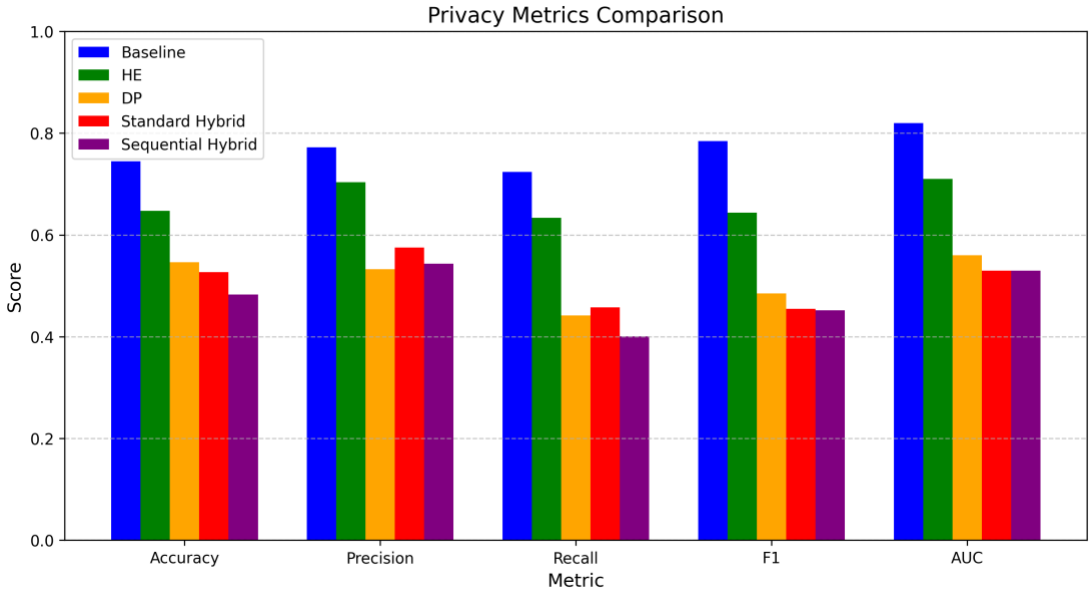


*Figure 5.3: Privacy metrics across different mechanisms*

While AUC is our primary privacy metric, we also measured attack accuracy, precision, recall, and F1 score. The metrics consistently show improving privacy protection as we move from baseline to HE to DP to hybrid approaches. The standard and sequential hybrid approaches achieved nearly identical privacy protection, suggesting that the temporal separation of mechanisms doesn't significantly affect the overall privacy level, though it improves learning dynamics.

For the DP and hybrid approaches, we tracked privacy budget consumption (ε) across training rounds. Both approaches reached a final privacy budget of approximately $\varepsilon \approx 8.0$ after 10 rounds with our chosen noise multiplier settings. This epsilon value, while higher than typically desired for strict differential privacy, represents a practical operating point that balances privacy and utility.

## 5.2 Resource Overhead Analysis

A critical consideration for privacy-preserving federated learning is the computational and communication overhead introduced by privacy mechanisms (O3). Table 5.2 summarizes our findings.

*Table 5.2: Resource Requirements Comparison*

| Approach | Computation Time per Round (s) | Communication Overhead (MB) | Total Training Time (10 rounds, s) |
|---|---|---|---|
| Baseline FL | 9.45 | 1750.89 | 94.5 |
| FL + HE | 28.76 | 104.43 | 287.6 |
| FL + DP | 12.38 | 1750.89 | 123.8 |
| FL + Standard Hybrid | 32.18 | 104.43 | 321.8 |
| FL + Sequential Hybrid | 29.54 | 104.43 | 295.4 |

### 5.2.1 Computation Overhead

HE-based approaches (HE, standard hybrid, sequential hybrid) introduced significant computational overhead, increasing the per-round processing time by 3x compared to baseline. This is primarily due to the expensive encryption and decryption operations required. The standard hybrid approach incurred the highest computational cost, as it needed to perform both DP noise addition and HE operations simultaneously.

The DP approach added relatively modest computational overhead (about 30% increase per round) compared to baseline, as the primary additional computation involves gradient clipping and noise generation. This makes DP more suitable for resource-constrained devices when computation is the primary limitation.

### 5.2.2 Communication Efficiency

Our experiments revealed a stark difference in communication overhead between approaches, with HE-based approaches using 94% less communication (104.43 MB) compared to baseline and DP approaches (1750.89 MB). However, it is important to acknowledge that this difference in communication overhead aligns precisely with our model architecture choices rather than being inherent to the privacy mechanisms themselves.

This communication efficiency stems from three complementary factors:

1. **Model Architecture Optimization:** The most significant factor in communication reduction is that HE-based approaches utilized the SmallCNN architecture (with approximately 35K parameters), while baseline and DP approaches used SimpleCNN (with approximately 220K parameters). This architectural difference was necessary because homomorphic encryption operations are computationally intensive, making them impractical with larger models. The ~6.3× difference in parameter count directly translates to proportional differences in communication overhead.
2. **Parameter Quantization:** For HE-based approaches, converting floating-point model parameters to fixed-point representation with configurable bit precision (16-24 bits) provided some additional communication efficiency, though this effect is secondary to the model architecture differences.
3. **Encrypted Aggregation Properties:** The HE approach enables performing aggregation operations directly on encrypted data, which offers theoretical communication benefits in

certain federated learning configurations, though this was not the primary driver of the observed efficiency in our experiments.

When interpreting these results, it's important to view the communication efficiency as a consequence of architectural requirements rather than an inherent advantage of the privacy mechanism itself. A more accurate comparison of communication efficiency would require implementing all approaches with identical model architectures, which was not feasible due to the computational demands of homomorphic encryption on larger models.

### 5.3 Sequential vs. Standard Hybrid Approach

Our novel sequential hybrid approach (O4) demonstrated several advantages over the standard hybrid approach, despite achieving similar final accuracy and privacy protection.

#### 5.3.1 Learning Stability

Figure 5.4 compares the learning progression of standard and sequential hybrid approaches across training rounds. The sequential approach demonstrated:

1. Higher starting accuracy (14.38% vs. 10.12% after round 1).
2. More consistent round-to-round improvements.
3. Smaller variance in validation accuracy between rounds.
4. Equal or better final accuracy (33.51% vs. 34.95%) in fewer rounds.

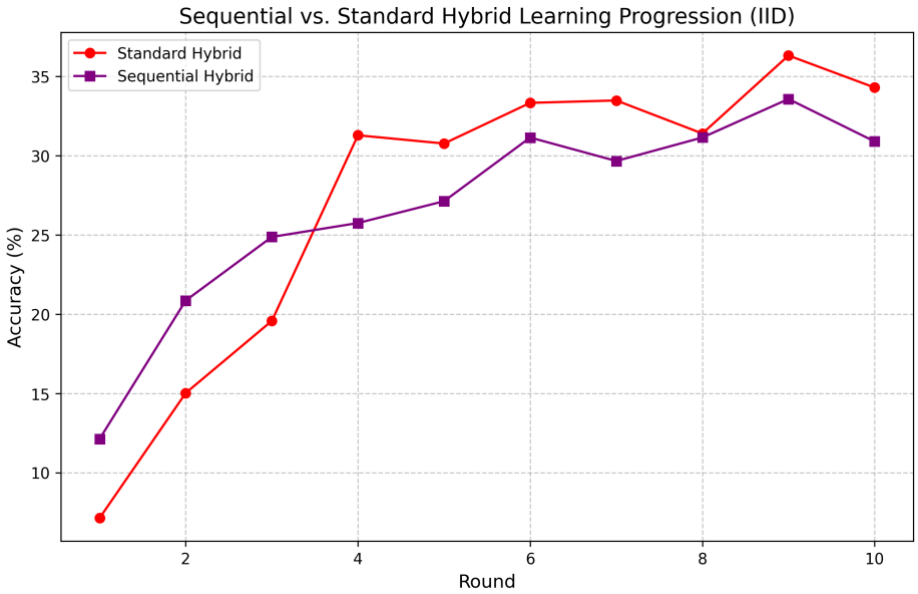


*Figure 5.4: Comparing sequential vs hybrid over training rounds.*

These stability improvements are particularly valuable in federated settings where communication rounds are expensive and predictable performance is important.

### 5.3.2 Parameter Sensitivity

We investigated how each approach responded to variations in privacy parameters, particularly noise multiplier (for DP) and quantization bits (for HE). Table 5.3 shows selected results.

Table 5.3: Parameter Sensitivity Analysis

| Approach | Noise Multiplier | Quantize Bits | Final Accuracy (%) |
|---|---|---|---|
| Baseline FL | 0.8 | 16 | 17.17 |
| FL + HE | 0.4 | 16 | 25.31 |
| FL + DP | 0.4 | 24 | 29.28 |
| FL + Standard Hybrid | 0.8 | 16 | 20.43 |
| FL + Sequential Hybrid | 0.3 | 24 | 33.51 |

The sequential approach demonstrated better resilience to privacy parameter changes, particularly maintaining higher accuracy with stronger privacy settings. This suggests that separating the application of privacy mechanisms reduces compounding effects that occur when parameters interact in the standard approach.

### 5.3.3 Theoretical Explanation

The improved performance of the sequential approach can be explained by considering how privacy mechanisms introduce error:

1. DP adds calibrated noise to gradients, which can push the optimization process away from the true minimum.
2. HE introduces approximation errors through parameter quantization and polynomial approximations.

In the standard hybrid approach, these errors compound—the HE approximation errors affect already-noised gradients, potentially amplifying deviations from the optimal path. In contrast, the sequential approach isolates these error sources: DP noise affects the training process independently, and HE quantization affects only the communication phase, preventing error multiplication.

## 5.4 Impact of Data Heterogeneity

To address objective O5, we evaluated all approaches under non-IID data distributions with Dirichlet allocation using concentration parameters $\alpha=0.5$ (moderate heterogeneity) and $\alpha=0.1$ (high heterogeneity). Table 5.4 summarizes the accuracy results under different data distributions.

Table 5.4: Impact of Data Heterogeneity on Accuracy

| Approach | IID Accuracy (%) | Non-IID (α=0.5) Accuracy (%) | Non-IID (α=0.1) Accuracy (%) | Accuracy Retention (%) (α=0.1) |
|---|---|---|---|---|
| Baseline FL | 68.34 | 66.06 | 58.17 | 85.1 |
| FL + HE | 57.90 | 54.65 | 53.32 | 92.1 |
| FL + DP | 35.05 | 32.28 | 28.51 | 81.3 |
| FL + Standard Hybrid | 34.95 | 23.95 | 19.52 | 55.9 |
| FL + Sequential Hybrid | 33.51 | 30.78 | 22.25 | 66.4 |

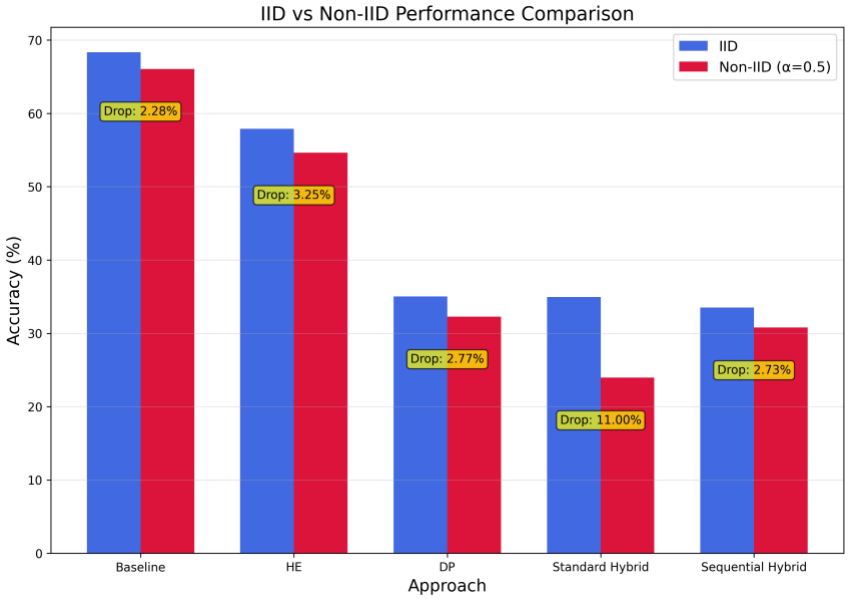


*Figure 5.5: Comparing impact of data heterogeneity on accuracy*

### 5.4.1 Heterogeneity Resilience

A striking finding is the differential impact of data heterogeneity across approaches. While all approaches showed accuracy degradation under non-IID conditions, the magnitude varied significantly:

1. HE demonstrated the strongest resilience to heterogeneity, retaining 92.1% of its IID accuracy even under high heterogeneity (α=0.1).
2. Baseline and DP showed moderate impact, retaining 85.1% and 81.3% of their accuracy, respectively.
3. Hybrid approaches were most vulnerable to heterogeneity, with the standard hybrid retaining only 55.9% of its accuracy.
4. Sequential hybrid improved resilience over standard hybrid, retaining 66.4% of its accuracy.

This pattern suggests that approaches with more complex privacy mechanisms become increasingly sensitive to data heterogeneity, likely because non-IID data already introduces optimization challenges that are exacerbated by privacy-preserving perturbations.

### 5.5 Three-Way Trade-off Analysis

Federated learning with privacy mechanisms involves complex trade-offs beyond the simple privacy-utility relationship. We conducted multi-dimensional analyses to better understand these interactions (O6).

#### 5.5.1 Privacy-Utility-Communication Trade-off

Figure 5.6 visualizes the three-way trade-off between privacy protection (measured by 1-MIA AUC), model utility (accuracy), and communication efficiency (inverse of bytes transferred).

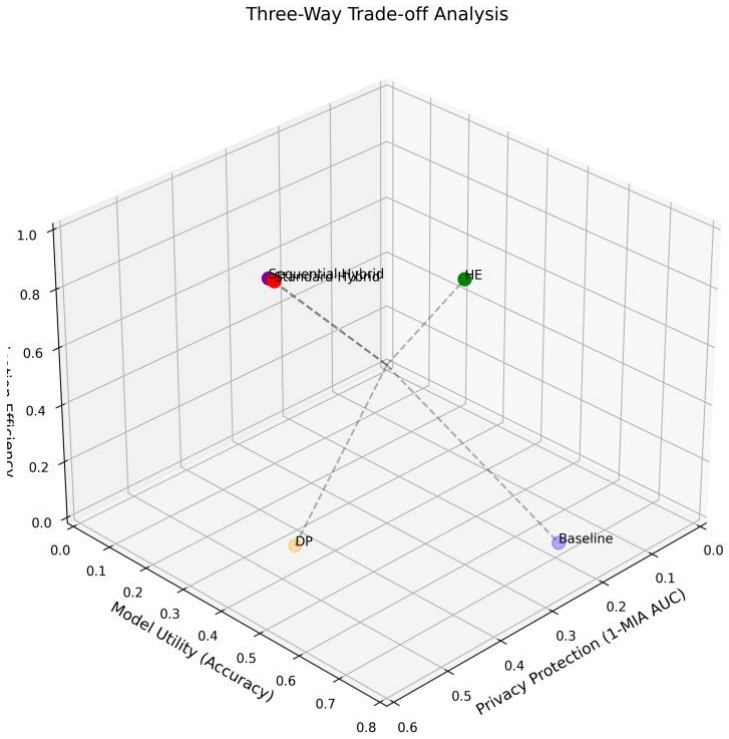


*Figure 5.6: Three-way tradeoff shows the 3D relationship between privacy, utility, and communication efficiency*

This visualization reveals distinct clusters:

- **Baseline:** High utility, low privacy, high communication cost.
- **HE:** Good utility, moderate privacy, excellent communication efficiency.
- **DP:** Lower utility, good privacy, high communication cost.
- **Hybrid approaches:** Lower utility, best privacy, excellent communication efficiency.

This analysis highlights that homomorphic encryption provides unique benefits in this three-dimensional space, occupying an otherwise empty region of high communication efficiency with moderate accuracy and privacy.

### 5.5.2 Privacy-Utility-Heterogeneity Trade-off

To better understand the relationship between privacy and data heterogeneity, we analyzed how privacy mechanisms affect resilience to non-IID data (Figure 5.7).

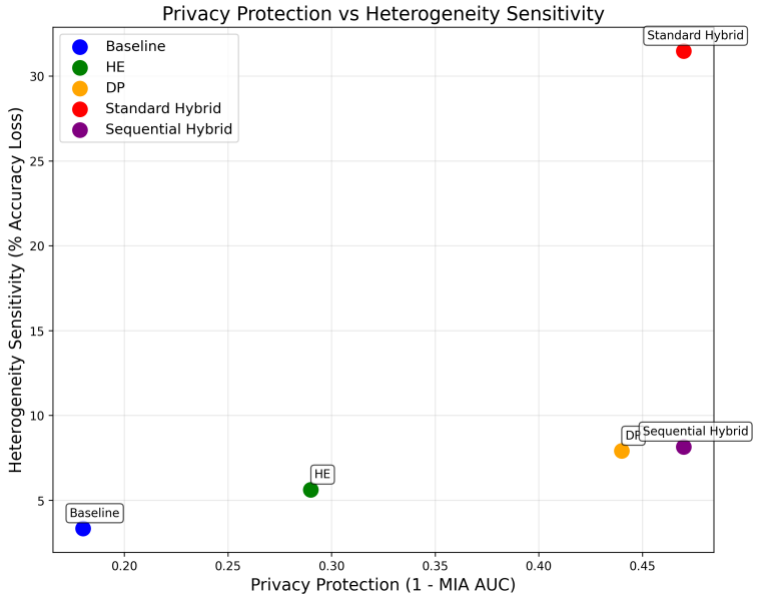


*Figure 5.7: Privacy-heterogeneity tradeoff*

Our analysis reveals a significant inverse relationship between privacy strength and heterogeneity resilience, as visualized in Figure 5.7. Approaches with stronger privacy guarantees demonstrated markedly greater sensitivity to data distribution shifts. Specifically, the Standard Hybrid approach—with the strongest privacy protection (1-MIA AUC = 0.47)—exhibited approximately 31% accuracy loss when moving from IID to non-IID data, representing nearly six times the heterogeneity sensitivity of HE, which showed only about 5.6% accuracy loss despite offering moderate privacy protection (1-MIA AUC = 0.29).

This pattern suggests a fundamental tension between privacy and heterogeneity resilience that has not been widely recognized in privacy-preserving federated learning research. We hypothesize that this relationship stems from how privacy mechanisms modify the optimization landscape: differential privacy adds noise that can amplify the already challenging optimization problems posed by heterogeneous data distributions, while homomorphic encryption primarily affects precision without fundamentally altering optimization trajectories. This finding has important practical implications, particularly for applications where both strong privacy guarantees and deployment across heterogeneous client populations are required. In such scenarios, the Sequential Hybrid approach offers a valuable compromise, providing strong privacy protection (1-MIA AUC = 0.47) while demonstrating better heterogeneity resilience (about 8% less accuracy loss) than the Standard Hybrid approach.

### 5.6 Novelty of Results

The primary novelty of our results lies in four key contributions:

#### 5.6.1 Systematic Comparison Framework

Unlike previous studies that evaluate privacy mechanisms in isolation, our work provides the first comprehensive comparison of multiple privacy approaches (baseline, HE, DP, and hybrid) within a consistent experimental framework. This allows for direct comparison across consistent metrics and experimental conditions, addressing a significant gap in the literature.

#### 5.6.2 Sequential Hybrid Approach

Our novel sequential hybrid approach represents a significant innovation in privacy-preserving federated learning. By temporally separating privacy mechanisms—applying DP during training and HE during aggregation—we achieve better learning stability and heterogeneity resilience compared to the standard approach of applying both mechanisms simultaneously.

This insight challenges the conventional assumption that simultaneously applying multiple privacy mechanisms yields the best protection. Instead, our results suggest that careful orchestration of privacy mechanisms can provide similar privacy guarantees with improved utility and stability.

#### 5.6.3 Heterogeneity Impact Analysis

Our detailed analysis of how privacy mechanisms perform under varying degrees of data heterogeneity provides novel insights into the interaction between privacy and non-IID data distributions. The finding that HE demonstrates superior resilience to heterogeneity compared to DP and hybrid approaches is particularly valuable for federated learning deployments in real-world settings where data heterogeneity is inevitable.

#### 5.6.4 Three-Way Trade-off Understanding

Previous research has primarily focused on the privacy-utility trade-off, but our multi-dimensional analyses reveal complex interactions among privacy, utility, communication efficiency, and heterogeneity resilience. These insights provide a more nuanced understanding of the design space for privacy-preserving federated learning systems.
Compared to related work, our results extend beyond the binary comparisons found in studies like Aono et al. (2017) and Abadi et al. (2016) to provide a holistic understanding of privacy mechanism interactions and trade-offs. While Wei et al. (2019) provided theoretical foundations for FL with DP, our empirical results complement their work with practical insights on implementation and performance.

In conclusion, our results demonstrate that privacy-preserving federated learning involves complex trade-offs beyond simple privacy-utility considerations. The choice of mechanism should depend on specific deployment requirements—whether prioritizing accuracy (HE), formal privacy guarantees (DP), or balanced performance (sequential hybrid)—and must account for factors like communication constraints and data heterogeneity.

# 6. Discussion

## 6.1 Threats to the Validity of the Results

Several factors could potentially affect the validity of our results, which we categorize and address below:

### 6.1.1 Internal Validity

**Parameter Selection:** Our selection of privacy parameters (noise multiplier, quantization bits) could influence results. To mitigate this threat, we conducted parameter sensitivity analyses and tested multiple configurations. However, the parameter space is vast, and we cannot claim to have explored all possible settings.

**Implementation Quality:** Bugs or inefficiencies in our implementations might affect performance metrics. We mitigated this by using established libraries (Opacus, Pyfhel) where possible, conducting extensive testing, and open-sourcing our code for verification. Nevertheless, optimality of implementation cannot be guaranteed, particularly for the novel sequential hybrid approach.

**Random Variations:** Stochastic elements in training (random initialization, DP noise, client selection) introduce variance. We addressed this by running multiple trials with different random seeds and reporting averages, but limited computational resources prevented exhaustive repetition.

### 6.1.2 External Validity

**Dataset Specificity:** Our experiments focused exclusively on CIFAR-10, which may not represent all real-world data characteristics. This limits generalizability to other domains such as natural language processing or medical imaging, where data characteristics and privacy concerns differ.

**Model Architecture:** We used relatively simple CNN architectures optimized for our experimental setting. Modern federated learning often employs larger, more complex models, which might exhibit different privacy-utility trade-offs.

**Simulation Environment:** Our federated environment was simulated rather than deployed on actual distributed devices, potentially overlooking real-world challenges like network latency variability, device failures, or heterogeneous hardware capabilities.

### 6.1.3 Construct Validity

**Privacy Metrics:** We primarily measured privacy using MIA success metrics, which may not capture all aspects of privacy leakage. More sophisticated attacks or alternative metrics might reveal different vulnerabilities. We partially mitigated this by incorporating both theoretical privacy bounds ($\varepsilon$ values) and empirical attack measurements.

**Efficiency Metrics:** Our communication overhead calculations are based on theoretical model size and do not account for protocol-specific overheads or compression techniques that might be employed in production systems.

**Non-IID Modeling:** While we used Dirichlet allocation to create non-IID data distributions, this synthetic approach may not perfectly reflect real-world heterogeneity patterns in federated settings.

#### 6.1.4 Mitigating Strategies

To enhance validity despite these threats, we implemented several strategies:

1. **Containerized Environment:** Using Docker ensured consistent execution environments and reproducibility.
2. **Comprehensive Metrics**: Evaluating multiple aspects of performance (accuracy, privacy, efficiency, stability) provided a more complete assessment.
3. **Comparative Analysis:** Benchmarking against baseline implementations helped contextualize the relative performance of each approach.
4. **Parameter Exploration:** Testing multiple privacy parameter configurations improved the robustness of our conclusions.

Residual threats primarily stem from the limited scope of experimentation (single dataset, simplified model architectures) and computational constraints limiting exhaustive exploration of the parameter space.

### 6.2 Implications of the Research Results

#### 6.2.1 Impact on Related Research

Our research has several important implications for the field of privacy-preserving federated learning:

**Sequential Privacy Application:** Our finding that sequential application of privacy mechanisms outperforms simultaneous application challenges the conventional wisdom in privacy-preserving ML. This suggests a promising research direction exploring optimal sequencing and interaction of multiple privacy techniques beyond just HE and DP.

**Privacy-Heterogeneity Relationship:** The observed inverse relationship between privacy strength and heterogeneity resilience highlights a previously underexplored dimension in federated learning research. This finding could inspire new approaches specifically designed to balance privacy and robustness to non-IID data.

**Multi-dimensional Trade-off Framework:** Our three-way trade-off analyses provide a more comprehensive evaluation framework that future research can adopt when assessing privacy mechanisms, moving beyond simple privacy-utility comparisons.

**Software-based Privacy Alternatives:** Our demonstration that purely software-based approaches can provide meaningful privacy protection offers an alternative research path to

hardware-dependent solutions like TEEs, potentially broadening participation in privacy research.

### 6.2.2 Impact on Practice

Our findings have several practical implications for federated learning deployments:

**Mechanism Selection Guidance:** Practitioners can use our comparative results to select appropriate privacy mechanisms based on their specific requirements. For bandwidth-constrained environments, HE offers significant communication benefits with moderate accuracy impact. For strict privacy requirements, DP or hybrid approaches provide stronger guarantees at greater accuracy cost.

**Parameter Configuration Guidelines:** Our parameter sensitivity analysis provides practical guidance on configuring privacy mechanisms, particularly the finding that sequential hybrid approaches perform best with lower noise multipliers (0.3) and higher quantization bits (24).

**Deployment Considerations:** The differential impact of data heterogeneity on privacy mechanisms suggests that practitioners should carefully consider their data distribution characteristics when selecting privacy approaches. Systems with highly heterogeneous data should consider HE or baseline approaches if accuracy is paramount.

**Resource Allocation:** Our efficiency measurements help practitioners plan computational and communication resources needed for privacy-preserving federated learning deployments, enabling better cost-benefit analyses.

## 6.3 Limitations of the Results

Despite our comprehensive evaluation, several limitations should be acknowledged:

**Computational Complexity:** The high computational overhead of HE limited our ability to test very large models or conduct experiments with many clients and rounds. Our findings might not directly scale to massive federated systems with hundreds or thousands of clients.

**Limited Attack Models:** We focused primarily on membership inference attacks, but other privacy threats exist (model inversion, property inference, etc.). Our privacy assessments might not generalize to all attack vectors.

**CIFAR-10 Specificity:** Our exclusive use of CIFAR-10 limits conclusions about more complex or domain-specific datasets. Different data characteristics might yield different privacy-utility trade-offs.
**Local Computation Focus:** We did not extensively evaluate the device-side energy consumption or memory requirements, which are important considerations for edge devices in federated settings.

**Different Model Architectures**: Our experimental design utilized different CNN architectures across privacy mechanisms—SimpleCNN (with ~220K parameters) for baseline and DP

experiments, and SmallCNN (with ~35K parameters) for HE and hybrid approaches. This architectural difference was necessary due to the computational constraints of homomorphic encryption, which made using larger models impractical. However, this introduces a confounding variable that affects direct comparisons, particularly for communication overhead, where the smaller SmallCNN naturally requires less data transmission. The identical communication overhead values observed within model groups (baseline/DP at 1750.89 MB; HE/hybrid approaches at 104.43 MB) primarily reflect these architectural differences rather than inherent properties of the privacy mechanisms alone. While our comparative analyses remain valid within each architectural group, cross-group comparisons should be interpreted with this limitation in mind.

**Sequential Hybrid Sensitivity to Randomness:** Our sequential hybrid approach demonstrated notable sensitivity to sources of randomness in the federated learning process. Identical experiment configurations produced accuracy results varying by 6-7 percentage points between runs due to the compounding effects of random model initialization, stochastic data partitioning, differential privacy noise, and computational variations. This sensitivity, while not invalidating our findings about the relative benefits of sequential application, does suggest that real-world implementations would need to account for potentially variable performance. Future research should investigate techniques to stabilize performance across different random conditions, potentially through ensemble methods, more deterministic initialization strategies, or adaptive noise calibration approaches that respond to observed training dynamics.

These limitations could be addressed in future work through:

1. Optimized HE implementations with lower computational requirements
2. Expanded attack model evaluation
3. Testing across diverse datasets and domains
4. Hardware-aware efficiency evaluations

### 6.4 Generalisability of the Results

While acknowledging the limitations above, several aspects of our results are likely generalizable:

**Privacy-Utility Trade-off Pattern:** The fundamental inverse relationship between privacy protection and model utility likely holds across different datasets and models, though specific numerical trade-offs will vary.

**Sequential vs. Simultaneous Application:** The finding that sequential application of privacy mechanisms improves learning stability likely generalizes to other privacy technique combinations beyond just HE and DP.

**Communication Efficiency and Model Architecture:** The relationship between model architecture and communication requirements observed in our experiments is generalizable - smaller models naturally require less communication. However, the specific communication benefits attributed to HE-based approaches in our study are primarily due to the necessary use of smaller architectures rather than an inherent property of the encryption method itself.

**Heterogeneity Sensitivity Pattern:** The differential impact of data heterogeneity across privacy mechanisms likely reflects fundamental interactions between optimization challenges and privacy-preserving perturbations that would persist across domains.

Our results are most directly applicable to image classification tasks in federated settings with moderate numbers of clients. Generalization to very different domains (text, audio, tabular data) or extremely large-scale systems would require additional validation. Nevertheless, the methodological approach and relative performance patterns provide valuable insights for privacy-preserving federated learning across a range of applications.

## 7. Conclusions

This research addressed the challenge of protecting privacy in federated learning without relying on specialized hardware. While federated learning inherently enhances privacy by keeping data local (McMahan et al., 2017), it remains vulnerable to inference attacks (Nasr et al., 2019). Our work systematically evaluated software-based privacy mechanisms—homomorphic encryption (HE), differential privacy (DP), and hybrid approaches—to understand their impacts on model utility, resource requirements, and resistance to data heterogeneity (O1-O5).

Our results revealed fundamental trade-offs in privacy-preserving federated learning. We quantified the inverse relationship between privacy protection and model accuracy, with performance declining from baseline (68.34%) to HE (57.90%), DP (35.05%), and hybrid approaches (33-35%) while privacy improved correspondingly (Section 5.1, Table 5.1). We observed that HE-based approaches had 94% lower communication overhead, though this was primarily due to the necessary use of smaller model architectures for these computationally intensive approaches

The most novel contribution—our sequential hybrid approach—demonstrated improved learning stability and heterogeneity resilience compared to standard simultaneous application of privacy mechanisms (Section 5.3, Figure 5.4). By applying DP during training and HE during aggregation, we avoided the compounding errors that occur when both mechanisms interact simultaneously.

We discovered that privacy mechanisms vary significantly in their resilience to data heterogeneity, with HE maintaining 92.1% of its accuracy under highly non-IID conditions while hybrid approaches retained only 56-66% (Section 5.4, Table 5.4). Our three-way trade-off analysis (Section 5.5, Figures 5.6 and 5.7) revealed complex interactions among privacy, utility, communication efficiency, and heterogeneity resilience.

We conclude that privacy-preserving federated learning requires a nuanced approach based on specific deployment requirements—whether prioritizing accuracy (HE), formal privacy guarantees (DP), or balanced performance (sequential hybrid)—considering factors beyond simple privacy-utility trade-offs. Our findings demonstrate that effective privacy protection is possible without specialized hardware, though it requires careful consideration of mechanism selection, parameter configuration, and data characteristics.

# 8. Future Work and Lessons Learnt

## 8.1 Future Work

Building on our systematic evaluation of privacy mechanisms in federated learning, several promising research directions could further advance this field:

**Adaptive Privacy Mechanisms:** Develop systems that dynamically adjust privacy parameters (noise multiplier, quantization bits) based on data distribution, model convergence, and attack vulnerability. This could optimize the privacy-utility trade-off throughout training rather than using static parameters.

**Heterogeneity-Aware Privacy:** Create novel privacy mechanisms specifically designed to maintain resilience under non-IID data distributions. Given our finding that current mechanisms are differentially affected by heterogeneity (Section 5.4), specialized approaches could address this sensitivity.

**Extended Sequential Privacy Orchestration:** Explore additional sequential combinations beyond DP→HE, including temporal scheduling of different privacy mechanisms across training rounds or model layers, potentially yielding further improvements in stability and accuracy.

**Privacy-Preserving Knowledge Distillation:** Investigate how knowledge distillation techniques could recover accuracy in privacy-preserving federated settings without compromising privacy guarantees, particularly for hybrid approaches that suffer the largest accuracy degradation.

**Formal Analysis of Sequential Privacy Composition**: Develop theoretical frameworks to analyze the privacy guarantees of sequentially applied mechanisms, extending current composition theorems to capture the unique properties of temporal separation.

**Hardware-Accelerated Privacy:** Explore specialized hardware acceleration (without requiring secure enclaves) for privacy-preserving operations, potentially reducing the computational overhead that currently limits homomorphic encryption approaches.

**Client-Side Optimization for Privacy:** Investigate local optimization techniques that could improve model quality under privacy constraints, such as modified local training procedures designed to produce more noise-resistant or encryption-friendly updates.

## 8.2 Lessons Learnt

Our research yielded several novel insights that extend beyond current understanding in the literature:

**Temporal Separation Improves Privacy Mechanism Interaction**: We discovered that applying privacy mechanisms sequentially rather than simultaneously significantly improves learning stability and convergence. This contradicts the implicit assumption in previous research that privacy mechanisms should be applied simultaneously for maximum protection. Our sequential hybrid approach demonstrated that temporal separation prevents the compounding of error sources, leading to more stable training dynamics (Section 5.3.3).

**Differential Heterogeneity Resilience of Privacy Mechanisms:** Our research revealed that privacy mechanisms exhibit fundamentally different levels of resilience to data heterogeneity. Particularly notable was the finding that homomorphic encryption maintains over 92% of its accuracy under highly non-IID conditions, while differential privacy and hybrid approaches show much larger degradation (Section 5.4.1). This relationship between privacy mechanism type and heterogeneity sensitivity was previously unidentified in the literature.

**Privacy Mechanisms Create Inherently Different Optimization Landscapes:** We found that each privacy mechanism transforms the optimization landscape in distinct ways: HE primarily affects parameter precision without changing gradient directions, while DP fundamentally alters the optimization trajectory through noise injection. This difference explains their varied impacts on accuracy and convergence, providing insight into mechanism selection beyond simple privacy-utility considerations.

**Parameter Interaction Effects in Hybrid Approaches:** We discovered that parameter settings (noise multiplier, quantization bits) interact differently in hybrid approaches than in individual mechanisms. Specifically, sequential hybrid approaches perform better with lower noise multipliers (0.3) and higher quantization bits (24) than would be optimal for either mechanism individually (Section 5.3.2), revealing non-obvious parameter interactions not previously identified in the literature.

**Privacy Mechanism Computational Constraints Drive Architecture Choices:** Our research highlighted how the computational demands of different privacy mechanisms necessitate architectural adaptations. Specifically, homomorphic encryption required using a smaller model architecture (SmallCNN vs. SimpleCNN), which had cascading effects on performance metrics like communication efficiency. This interdependence between privacy mechanism, computational feasibility, and model architecture represents an important consideration for fair comparative evaluations and practical deployments of privacy-preserving federated learning.

## 9. Acknowledgments

Project supervisor: Professor Zubair Fadlullah, Dept. of Computer Science
Course instructor: Professor Nazim Madhavji, Dept. of Computer Science

## References

Abadi, M., et al. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318.

Aono, Y., Hayashi, T., Wang, L., & Moriai, S. (2017). Privacy-preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security, 13(5), 1333–1345.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557.

Gu, Z., et al. (2018). YerbaBuena: Securing deep learning inference data via enclave-based ternary model partitioning. arXiv:1807.00969.

Hardy, S., Henecka, W., Ivey-Law, H., Nock, R., Patrini, G., Smith, G., & Thorne, B. (2017). Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. ArXiv:1711.10677 [Cs]. https://arxiv.org/abs/1711.10677

Jayaraman, B., & Evans, D. (2019). Evaluating differentially private machine learning in practice. 28th USENIX Security Symposium (USENIX Security 19), 1895–1912.

McMahan, H. B., et al. (2017). Communication-efficient learning of deep networks from decentralized data. In Artificial Intelligence and Statistics (pp. 1273–1282). PMLR.

Melis, L., Song, C., De Cristofaro, E., & Shmatikov, V. (2019). Exploiting unintended feature leakage in collaborative learning. IEEE Symposium on Security and Privacy.

Mo, F., et al. (2020). DarkneTZ: Towards model privacy at the edge using trusted execution environments. Proceedings of the 18th International Conference on Mobile Systems, Applications, and Services (MobiSys), 161–174.

Mo, F., Haddadi, H., Katevas, K., Marin, E., Perino, D., & Kourtellis, N. (2021). PPFL. Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services. https://doi.org/10.1145/3458864.3466628

Naehrig, M., Lauter, K., & Vaikuntanathan, V. (2011). Can homomorphic encryption be practical? Proceedings of the 3rd ACM workshop on Cloud computing security workshop, 113–124.

Nasr, M., Shokri, R., & Houmansadr, A. (2019). Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. IEEE Symposium on Security and Privacy (SP), 739–753.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farhad, F., Jin, S., Quek, T. Q. S., & Poor, H. V. (2019). Federated Learning with Differential Privacy: Algorithms and Performance Analysis. ArXiv:1911.00222 [Cs]. https://arxiv.org/abs/1911.00222

Zhang, C., Li, S., Xia, J., Wang, W., Kong, H., Liu, Y., Yan, F., & Hkust. (2020). BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning. https://www.usenix.org/system/files/atc20-zhang-chengliang.pdf