Privacy-Preserving Machine Learning in Practice

ANEESH PATEL, University of California, Berkeley, USA

The increasing prevalence of machine learning (ML) models underscores the critical need for robust safeguards to protect sensitive data, thus elevating privacy-preserving ML (PPML) as a pivotal area of research. This thesis delves into the practical intricacies of PPML by examining and evaluating the effectiveness of various techniques in real-world scenarios.

Driven by mounting concerns surrounding data breaches and privacy attacks, this research prioritizes privacy throughout the ML lifecycle. Recognizing the inherent risks posed by compromised data – including biased models, vulnerability exploitation, and privacy violations – we also acknowledge the broader societal implications, such as eroded trust, hindered applications, and exacerbated inequalities [61].

We examine three existing and one novel PPML approach:

- Homomorphic Encryption: Enables training on encrypted data, ensuring the data curator never has to see the raw data.
- Federated Learning: Facilitates decentralized training, allowing devices to collaboratively learn without sharing raw data.
- Differential Privacy: Provides mathematical guarantees on information leakage about data points used in model training.
- Memorizing Models for Synthetic Data Generation: Overfits a memorizing model to generate a private, synthetic, labeled dataset.

In this paper, we evaluate federated learning, differential privacy, the combination of the two, and memorizing models for synthetic data generation on benchmark datasets from real-world applications, scrutinizing their effectiveness in safeguarding privacy while preserving model performance and computational efficiency, as well as identifying their shortcomings and how to best address them. Ultimately, we devise a comprehensive framework for how PPML techniques can be applied practically, empowering practitioners, guiding future research, and encouraging the development of secure and privacy-compliant ML models.

Additional Key Words and Phrases: Privacy-Preserving Machine Learning, Privacy, Machine Learning, Homomorphic Encryption, Federated Learning, Differential Privacy

1 INTRODUCTION

Machine learning (ML) models have taken the world by storm. Propelled by recent advancements in artificial intelligence (AI), these models have revolutionized fields such as computer vision and natural language processing, while giving rise to entirely new ones, like conversational AI. Notably, the advent of generative AI has unlocked unprecedented capabilities, enabling the creation of realistic text, images, and video, further expanding the horizons of creative expression and automation [39].

However, amidst this remarkable advancement, a critical consideration emerges regarding the foundation upon which these models thrive: data. ML models, particularly deep learning architectures, exhibit an avid appetite for data, often consuming large volumes to train and refine their parameters. This heavy reliance on data, while instrumental for enhancing model accuracy and performance, raises important questions concerning data privacy and security [12]. As these models ingest and process massive datasets, they become custodians of potentially sensitive information, encompassing personal identifiers, demographic information, and intimate behavioral patterns. Consequently, the access to and utilization of personal data, both explicit and implicit, within ML ecosystems can pose a host of privacy risks, giving rise to multifaceted challenges in safeguarding individual privacy.

These challenges manifest in privacy breaches, stemming from unauthorized access to or unintentional disclosure of personal data. As a result, the confidentiality and integrity of individuals' information are compromised. Various

Author's address: Aneesh Patel, AneeshPatel@berkeley.com, University of California, Berkeley, Berkeley, CA, USA, 94704.

sophisticated attacks have emerged, including reconstruction attacks, linkage attacks, and membership inference attacks, in which adversaries exploit ML models to deduce sensitive information from seemingly innocuous data or queries with alarming precision [54].

In response to these challenges, regulatory frameworks such as the General Data Protection Regulation (GDPR) and the California Privacy Rights Act (CPRA) underscore the urgent nature of addressing privacy concerns in ML ecosystems [15]. These legal directives exemplify concerted efforts to mitigate privacy risks and proactively enforce ethical obligations incumbent upon organizations entrusted with sensitive data.

Against this backdrop of regulatory scrutiny and burgeoning privacy concerns, there exists a compelling need for innovative solutions that reconcile the value of data-driven insights with the imperative of safeguarding individual privacy. Privacy-preserving machine learning emerges as a pertinent paradigm, offering robust privacy-preserving mechanisms that preserve the efficacy and utility of ML models.

Recent years have witnessed a surge of interest and investment in PPML research, driven by the recognition of privacy vulnerabilities inherent within conventional ML frameworks. Despite this, a gap persists – a lack of an accessible framework for translating theoretical advancements in PPML into practical and deployable solutions.

To bridge the divide between theory and practice, this paper focuses on four pivotal privacy-preserving mechanisms: homomorphic encryption, federated learning, differential privacy, and memorizing models for synthetic data generation. We start with a discussion of the socio-technical motivations for privacy preservation, before providing a conceptual overview of the mathematical and theoretical underpinnings of these mechanisms. Building upon this conceptual framework, we then conduct a rigorous comparative analysis of these mechanisms, examining their respective benefits, trade-offs, and sample use cases. Transitioning to practical considerations, we then examine the performance of both publicly available and custom implementations of these mechanisms when applied to realistic datasets. The private and non-private models are evaluated against one another using metrics that evaluate accuracy, efficiency, and privacy.

Ultimately, this paper aims to serve as a comprehensive framework delineating best practices and guidelines to deploy PPML in real-world settings. By integrating theoretical insights with practical implementations, this paper offers actionable guidelines and a holistic blueprint for navigating the complex terrain of privacy preservation within the ML ecosystem.

2 LITERATURE REVIEW

This thesis draws upon privacy and security research from three key fields: differential privacy, homomorphic encryption, and federated learning. Each offers distinct advantages and challenges in securing sensitive data within the ML pipeline. Through a review of the existing literature, we analyze the applicability, limitations, and recent advancements in each field. However, before delving into these specialized domains, it is crucial to establish a firm foundation of understanding. We first take a brief detour into the fundamental concept of privacy itself, aiming to provide a shared perspective on this multifaceted challenge. Then, to contextualize the subsequent exploration, we examine some common attacks and vulnerabilities of ML models, looking at their conceptual workings and providing illustrative examples.

2.1 Concept of Privacy

Prior to implementing privacy measures, it is essential to grasp the fundamental concept of privacy, why it matters, and its implications within the realm of machine learning. By understanding these foundational aspects, one can effectively identify and address privacy concerns tailored to their specific use case.

2.1.1 What is Privacy? The notion of privacy, far from being a static concept, is a living entity that has evolved over time. It has been shaped by societal norms, cultural values, and technological advancements, and will continue to do so. Central to privacy is the innate human desire for autonomy and control over personal information [40]. From physical seclusion to digital data management, however, privacy has been interpreted in different ways throughout history [44].

While the absence of a universally accepted definition complicates discussions surrounding privacy, paradoxically, it is this very absence that is crucial to understanding the concept of privacy [50]. With different situations calling for vastly different approaches to privacy, it is imperative that privacy solutions are designed to be nuanced and context-specific.

To organize privacy analysis, Mulligan, Koopman, and Doty introduce the *privacy analytic*, a framework that maps privacy considerations across 14 dimensions, encompassing theory, protection, harm, provision, and scope [50]. Blumenstock and Kohli describe a light-weight version of the privacy analytic for practitioners in terms of the "who-what-when-where-why-how"'s of privacy threats, which focuses on the human-centric nature of privacy and its violations [16]. By adopting this approach, privacy analysis remains anchored in real-world contexts, ensuring a holistic understanding of privacy concerns and facilitating the development of tailored solutions.

2.1.2 Why does Privacy Matter? Privacy holds profound significance in modern society, serving as a cornerstone of individual autonomy, dignity, and freedom [40]. Solove explains privacy is crucial to preserve respect for others, maintain social boundaries, and allow people the ability to change, among others [59].

Privacy breaches and violations can have far-reaching implications, jeopardizing not only individuals' personal autonomy but also societal trust and stability [21]. When sensitive information falls into the wrong hands, it can be exploited for nefarious purposes, leading to identity theft, financial fraud, and emotional harm [62]. Moreover, privacy violations can exacerbate power imbalances and perpetuate social injustices, particularly when vulnerable communities are disproportionately affected [57].

In an era characterized by rapid technological advancement and ubiquitous data collection, the importance of protecting privacy cannot be overstated. As individuals and organizations navigate the digital landscape, it is essential to prioritize privacy considerations and implement robust safeguards to protect personal information from unauthorized access and misuse. By doing so, we can ensure that privacy remains a fundamental human right in an increasingly interconnected world.

2.1.3 Privacy in Machine Learning. In the Information Age, traditional privacy analysis typically encompasses four critical components: information collection, storage, processing, and dissemination [60]. Within the context of machine learning, these fundamental privacy considerations translate directly to data collection, data storage, model training, and model output. Ensuring meaningful privacy guarantees requires safeguarding privacy at each stage of the process; neglecting any aspect of this pipeline not only compromises the privacy of the individual component but also undermines the overall privacy integrity of the entire system and exposes it to potential privacy breaches and vulnerabilities.

Data Collection. The initial stage of the ML pipeline involves gathering diverse datasets from various sources, which raises privacy concerns due to the potential exposure of sensitive information during acquisition. This stage is inherently context-specific and varies widely across applications and domains, so we merely introduce this idea for completeness, and encourage organizations to conscientiously navigate the ethical and legal implications related to data minimization, consent management, and data provenance to ensure responsible data collection practices [18].

Data Storage. Securing data at rest is crucial to protect sensitive information from unauthorized access or disclosure. As organizations amass vast repositories of data, the complexity of maintaining data security escalates significantly, but so does the importance of doing so. It is imperative to safeguard sensitive data from unauthorized access, breaches, and misuse to mitigate the risks of privacy violations. Individuals entrusting their data to providers rightfully expect their information to be safeguarded in a secure manner [3]. Furthermore, privacy considerations extend to situations where data must remain private from even the data curators themselves [9]. Throughout the data storage life cycle, privacy entails not only safeguarding the data from attacks and breaches, but also potentially storing it in a form that is private from the curators themselves.

Model Training. During model training, the primary privacy concern relates to collaborative learning, wherein multiple parties collaborate to train a single ML model using their respective datasets. While exposing a model to more and potentially more diverse data points typically enhances model performance, there are scenarios where the data points represent sensitive information and cannot or should not be shared directly [42]. For instance, hospitals may seek to train a chest radiography classification model without disclosing patient mammograms, or a software company may aim to refine a smart keyboard suggestion model without accessing the raw text data from users [37] [65]. In these cases, privacy entails collaboratively training models without necessitating access to the raw data.

Model Output. The final stage involves querying the model to generate predictions or classifications. Privacy concerns arise if these outputs inadvertently reveal specific information about the training dataset. Even seemingly innocuous outputs can be exploited by malicious actors within just a few queries to divulge information about individuals in the training set [42]. Therefore, it is paramount to ensure model outputs prevent any information leakage about individual data points in the dataset to maintain the confidentiality and integrity of sensitive information.

2.2 Attacks and Vulnerabilities

As we transition from our exploration of the fundamental concept of privacy, we now turn our attention to specific threats and vulnerabilities that can compromise privacy within the ML pipeline, specifically delving into the nuances of attacks and vulnerabilities during data storage, model training, and model output.

2.2.1 Data Storage. Data storage poses significant privacy risks, particularly when sensitive information is stored in its raw form. Raw data is vulnerable to various types of privacy breaches and attacks, including unauthorized access, data breaches, and malicious exploitation [10]. Furthermore, in some scenarios, there may be a need to keep the data private from even the data curators themselves [9]. This dual threat model includes both an untrustworthy storage platform and an untrustworthy curator, where "untrustworthy" denotes a party that should not have access to the raw sensitive data. An example of this is the 2017 Equifax data breach, where cyber-criminals exploited a vulnerability in the organization's data storage infrastructure to access the sensitive information of up to 143 million individuals [45]. This included Social Security numbers, addresses, financial information, and more, leading to widespread identity theft and financial fraud.

2.2.2 Model Training. Model training, especially in collaborative learning scenarios, introduces vulnerabilities related to sharing raw data over insecure channels. When data is exchanged between collaborators without adequate security measures, it becomes susceptible to various types of attacks and privacy breaches, including interception and data manipulation. In some cases, there may be a need to keep the data private not only from external threats but also from

the collaborators themselves [41]. This expanded threat model encompasses both an untrustworthy sharing channel and untrustworthy collaborators.

2.2.3 Model Outputs. There are several well-defined attacks involving model outputs that pose significant privacy risks, a few of which are outlined below.[1]. These attacks share a common threat model centered around an untrustworthy client, where "client" refers to any entity capable of querying the model.

Reconstruction Attacks. Reconstruction attacks exploit information leakage in the model's outputs to reverse-engineer the sensitive training data, thereby compromising the privacy of the individuals represented in the dataset [20]. The primary goal is to recover sensitive attributes or patterns present in the original data. For example, Garfinkel, Abowd, and Martindale were able to demonstrate how aggregate Census data could be used to reconstruct the microdata of individuals with alarmingly high accuracy [30]. Recent research has also shown that black-box ML algorithms can be exploited to reconstruct accurate samples with strong performance [56].

Linkage Attacks. Linkage attacks aim to leverage auxiliary information to identify individuals from the model's outputs. By correlating predictions with external datasets or public information, adversaries can link individuals to their sensitive attributes. Narayanan and Shmatikov famously showed how individual users of an "anonymized" Netflix dataset could reliably be identified using publicly available auxiliary data from the Internet Movie Database (IMDb) [51]. ML models, particularly those that seek to be explainable, have also demonstrated vulnerability to a range of linkage attacks [31].

Membership Inference Attacks. In these attacks, adversaries attempt to determine whether a specific data point was present in the model's training dataset. Research indicates that membership inference attacks can be executed effectively without any prior knowledge of the training data distribution [58]. In fact, even when the attacker's assumptions about the training data are imprecise, these attacks remain potent [58]. Demonstrations of membership inference attacks have been conducted across various models and datasets, including a cancer classification dataset, where target patients were able to be identified with 88% precision [43].

Stage	*Privacy From-Whom*	*Main Vulnerabilities*
Data Storage	Untrustworthy Storage Platform, Curator	Privacy Breaches, Unauthorized Access
Model Training	Untrustworthy Sharing Channel, Collaborators	Privacy Breaches, Data Interception
Model Output	Untrustworthy Client	Reconstruction, Linkage, Membership Inference

Table 1. Privacy Vulnerabilities in the ML Pipeline

2.3 Privacy-Preserving Approaches

2.3.1 Homomorphic Encryption.

Motivation: Securing Data at Rest and in Use. In today's data-driven world, safeguarding sensitive information is paramount, whether it's stored or being processed. While traditional encryption methods effectively protect data at rest, they fall short in ensuring security when it is in use. During computation, raw data is vulnerable to interception and

[1]In fact, Article 29 WP 216 of the GDPR specifically lists linkage and inference attacks as two of the critical threats that data practitioners must guard against[2]

unauthorized access, posing significant risks to data confidentiality and integrity. Moreover, conventional computation on raw data also raises privacy concerns by granting data curators access to potentially sensitive information. Placing trust solely in the security of the infrastructures or in the individuals handling the data is no longer sufficient in mitigating these risks.

Homomorphic encryption offers a promising solution to address these vulnerabilities and enhance data security. Unlike traditional encryption methods, which require decryption before computation, homomorphic encryption enables computation directly on encrypted data without decryption. This approach maintains data confidentiality throughout the computation process and ensures privacy from both external threats and internal entities, mitigating the risks associated with exposing raw data during computation [9].

Conceptual Understanding: Encrypted Computation without Data Exposure. Homomorphic encryption is a cryptographic technique that allows computations to be performed on encrypted data without the need for decryption. At its core, homomorphic encryption ensures that operations performed on encrypted data produce results that are consistent with those obtained from performing the same operations on the raw data.

Fundamentally, homomorphic encryption achieves this by preserving the mathematical structure of the raw data within its encrypted form. This means that mathematical operations performed on the encrypted data yield results that remain encrypted, maintaining the confidentiality of the underlying information.

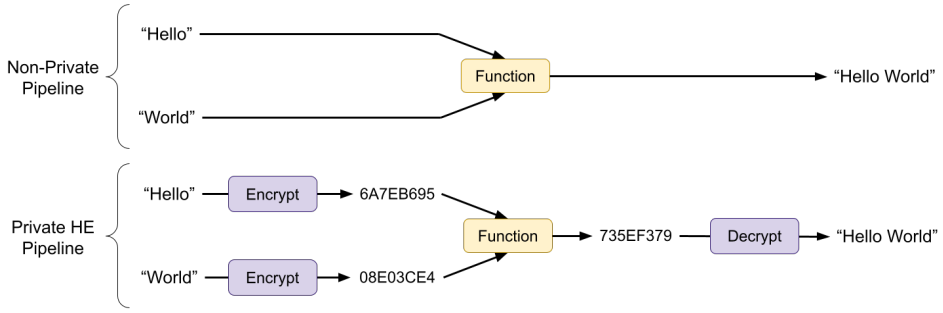


Fig. 1. Homomorphic Encryption Overview

Theoretical Underpinnings: Computing in a Ciphertext World. Homomorphic encryption is grounded in sophisticated mathematical principles that enable computations to be performed directly on encrypted data without requiring decryption. This cryptographic technique encompasses various encryption schemes categorized based on the operations they support and the extent of their capabilities [9]:

- *Partially Homomorphic Encryption (PHE)*: supports one type of operation an unlimited number of times (ex. RSA, Pallier)
- *Somewhat Homomorphic Encryption (SWHE)*: supports several types of operations, but these operations can only be performed a limited number of times (ex. BGN)
- *Fully Homomorphic Encryption (FHE)*: supports an unlimited number of operations, which can be performed an unlimited number of times (ex. BGV, BFV, CKKS)

In machine learning applications, where complex computations involving both addition and multiplication are commonplace, FHE is particularly relevant. By allowing computations to be conducted directly on encrypted data,

homomorphic encryption ensures that the underlying information remains confidential while still facilitating essential data processing tasks.

The security and effectiveness of homomorphic encryption depend heavily on the management of cryptographic keys. Keys play a critical role in securing encrypted data and controlling access to decryption processes. Key selection involves choosing appropriate parameters and configurations that align with the desired security and performance requirements of the encryption scheme. In fact, homormophic encryption schemes typically include a security parameter that determines the level of privacy guarantees and the robustness of the encryption against attacks. Securely distributing and managing keys among authorized parties is vital to ensure that only authorized users can encrypt and decrypt the raw data. Effective key management practices are paramount to the overall security and efficiency of homomorphic encryption systems, enabling robust protection of sensitive information in diverse use cases [9].

Tradeoffs: Privacy vs Computational Efficiency. Homomorphic encryption offers strong privacy guarantees; however, these benefits come with tradeoffs, particularly in terms of computational efficiency.

Analyzing these tradeoffs reveals that homomorphic encryption introduces significant computational overhead compared to traditional computation. Performing complex operations directly on encrypted data increases the demand for computational resources and processing time. Fully Homomorphic Encryption, which supports both addition and multiplication operations an arbitrary number of times, is particularly resource-intensive due to the complexity of its underlying mathematical operations. Additionally, higher levels of security come at the cost of increased overhead and reduced time performance [26].

The impact of homomorphic encryption on computational resources can result in increased latency and higher resource consumption. Encryption and decryption operations, especially in the context of FHE, are computationally expensive, requiring substantial processing power and memory.

Achieving a balance between privacy and computational efficiency is crucial for the practical adoption of homomorphic encryption in real-world applications. Optimizing implementations through algorithmic improvements and innovative approaches will be essential in harnessing the full potential of homomorphic encryption for secure and privacy-preserving data processing.

2.3.2 Federated Learning.

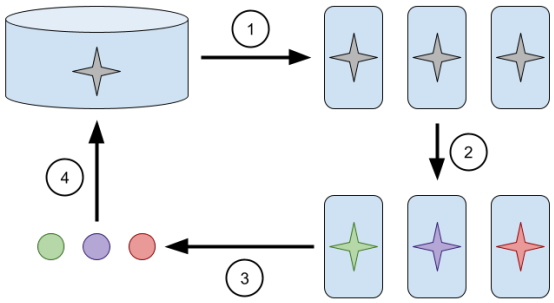
Motivation: Preserving Privacy in Collaborative Model Training. In collaborative model training scenarios, the preservation of privacy emerges as a critical concern. Traditional centralized approaches to model training necessitate the aggregation of data from multiple sources into a single repository, raising significant privacy risks. With sensitive data pooled together, there is heightened vulnerability to breaches and unauthorized access, potentially compromising the confidentiality of individuals' information. Consequently, there is a need for privacy-preserving techniques that can mitigate these risks while enabling effective collaboration and model improvement. Federated learning arises as a promising solution to address these challenges by allowing model training on decentralized data sources without compromising individual privacy [46].

Conceptual Understanding: Training in Decentralized Data Environments. Federated learning introduces a decentralized approach to model training, enabling the training process to occur across multiple data sources distributed throughout a network. This decentralization facilitates model training while respecting data privacy and security concerns inherent in centralized approaches.

In traditional centralized approaches to machine learning, data points are aggregated from various devices or servers across a network, and the model is trained on this aggregated data. The model continuously refines itself by updating its parameters, based on the information contained in each new data point. However, federated learning differs in that it aggregates these updates to the parameters, as opposed to the raw data points themselves [46].

This distinction offers significant advantages for data privacy preservation. By conducting model training locally on individual devices, federated learning minimizes the risks of unauthorized access or data breaches associated with centralizing data. Additionally, local model training ensures that sensitive data points remain within their respective environments, reducing exposure to external entities [67]. This distributed approach enhances privacy protection while facilitating collaborative model training across decentralized data sources.

Theoretical Underpinnings: Secure Aggregation and Model Updates. In federated learning, the conventional practice of aggregating all data in a single location for model training is replaced by a decentralized approach. Initially, a central model is distributed to individual devices or servers participating in the learning process. These devices, each holding sensitive data, train their local models using their respective datasets. The updates to the model's parameters are then transmitted back to the central server, where they are aggregated, typically using secure aggregation techniques, and applied to the central model [46]. This process, outlined in Figure 2, ensures that the central model learns from the collective knowledge of all participating devices, without directly accessing the raw data.



(1) Central server transmits central model to devices
(2) Devices train their models using local sensitive data
(3) Model updates are aggregated
(4) Aggregated model updates are sent to the central server and applied to the central model

Fig. 2. Federated Learning Overview

Secure aggregation techniques play a pivotal role in preserving privacy during federated learning. These techniques enable the central server to aggregate model updates from multiple devices or servers while maintaining the privacy of individual data contributions. Among the many ways of achieving secure aggregation during federated learning, differential privacy and homomorphic encryption are two that stand out [67]. Differential privacy helps to ensure that the model updates themselves prevent privacy leakage, and homomorphic encryption allows these updates to be transmitted and aggregated in an encrypted state without needing to decrypt them. Collectively, these methods ensure that the central server learns only the privatized aggregate updates to the model parameters, preventing any exposure of sensitive information during the aggregation process.

Tradeoffs: Privacy vs Computational Efficiency. Federated learning offers a paradigm shift in model training, yet it introduces a key tradeoff between privacy and computational efficiency that must be carefully considered. Decentralized training introduces significant computational overhead compared to traditional centralized approaches. The need to distribute model training across multiple devices or servers results in increased communication costs and computational complexity. While each local device trains a model independently, the coordination and synchronization of model updates across the network incur additional computational resources. As a result, while federated learning helps achieve privacy, it may exhibit slower convergence rates and higher resource consumption compared to centralized training methods [68].

This inherent tradeoff between privacy and computational efficiency necessitates careful optimization and resource allocation to strike an appropriate balance; addressing this requires a nuanced understanding of the specific application context.

2.3.3 Differential Privacy.

Motivation: Safeguard Individual Privacy. The motivation behind differential privacy stems from the increasing concern over individual privacy in the era of big data and machine learning. With the proliferation of data collection and analysis, there is a growing recognition of the need to protect sensitive information while still allowing for meaningful data analysis. Traditional privacy protection measures often fall short in the face of modern data analysis techniques, leaving individuals vulnerable to privacy breaches and misuse of their personal data.

Differential privacy provides strong and measurable privacy guarantees that protect individuals' sensitive information while allowing for useful data analysis. It achieves this by applying carefully curated statistical mechanisms to ensure that the presence or absence of any individual's data has a negligible effect on the overall output [64]. Consequently, even if an adversary has access to the output of a differentially private algorithm, they cannot infer whether any particular individual's data was included in the analysis with any reasonable degree of confidence. In the context of machine learning, this grants individuals a degree of plausible deniability regarding their presence within the training dataset [66].

One of the key aspects of differential privacy is that it provides a quantifiable metric for privacy loss [24]. This proves particularly useful when tracking the cumulative privacy loss across multiple differentially private analyses, which is especially useful in scenarios where ML models undergo multiple iterations over the training set [64].

Additionally, differential privacy provides the benefits of being both future-proof and adversary-agnostic [24]. Its future-proof nature ensures that even as data analysis techniques and computational capabilities evolve, its privacy guarantees remain robust. Furthermore, differentially privacy is adversary-agnostic, which means any further privacy loss is prevented regardless of the specific attack strategies employed by potential adversaries or any auxiliary information at their disposal.

Finally, the post-processing guarantee of differential privacy is a critical aspect of its framework, which stipulates that any function applied to the output of a differentially private mechanism will also preserve privacy, regardless of the specific function or subsequent analysis conducted [52]. This means that individuals' privacy remains safeguarded regardless of how the data is processed or utilized downstream.

Taken together, these attributes render differential privacy a highly versatile and broadly applicable framework for preserving privacy in ML model outputs across diverse domains.

Conceptual Understanding: Noise for Privacy Enhancement. Differential privacy operates on the principle of strategically introducing noise into data to obscure the individual contributions of specific data points.

Consider a scenario involving a dataset containing the income of individuals within a neighborhood. While the dataset is kept private, the average income is made public, and it is known there are 50 residents in the neighborhood. Without the protections of differential privacy, the published average income is $50,000.

Now, suppose a new individual moves into the neighborhood, and the average income jumps to $50,500. Using simple arithmetic, one could deduce the new individual's income as $75,500, which illustrates how without privacy protections, even seemingly innocuous aggregate statistics can inadvertently reveal sensitive individual information.

However, with the application of differential privacy, the published average income is now perturbed with carefully calibrated noise. Before the new individual's arrival, the published average income might be $51,276, increasing to $51,707 afterward. Attempting to perform the previous calculation to determine the new individual's income now yields $73,257. This inclusion of random noise ensures that the true value of the individual's income is protected.

In the context of machine learning, this noise injection can occur at various stages to uphold differential privacy guarantees:

(1) at the raw data level, generating a new, synthetic dataset
(2) during model training, applied to gradients or optimization processes
(3) to the model outputs themselves

Notably, due to the post-processing guarantee of differential privacy, noise introduction at any stage ensures privacy protection throughout all subsequent stages [52]. This distinction underscores that the choice of where noise is injected has nuanced implications for determining what aspects of data analysis can be considered public, as highlighted in Figure 3. While each approach serves to protect individual privacy, this paper primarily focuses on the incorporation of differential privacy during model training.

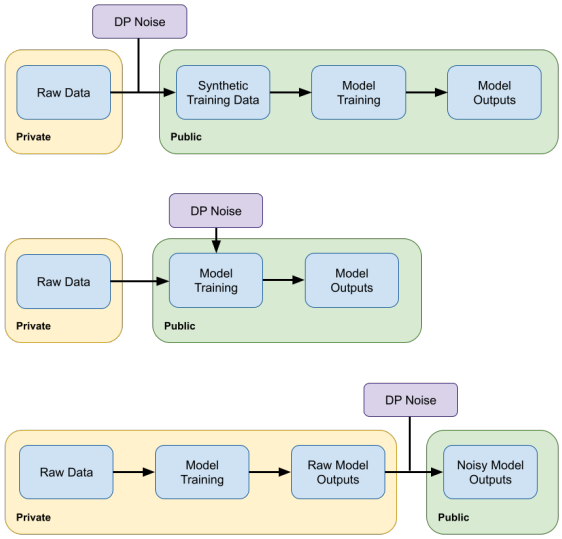


Fig. 3. Differential Privacy in Machine Learning Workflows

Theoretical Underpinnings: Privacy Budgets and Mechanisms. At the core of privacy-preserving mechanisms lies the notion of a privacy budget, which quantifies the maximum allowable privacy loss from any single computation or data release [64]. This budget represents a delicate balance between safeguarding privacy and preserving data utility, where a lower budget corresponds to a stricter level of privacy protection. Notably, the allocation of the privacy budget depends on the nature of the queries performed on the data. For instance, if only a single query is to be performed on the data, then that query can consume the entirety of the budget, whereas performing a series of queries requires that the budget is distributed across all the queries.

In the framework of differential privacy, the parameter epsilon (ϵ) represents the magnitude of the privacy budget allocated to a specific computation or data release. It quantifies the maximum allowable deviation in the output of a differentially private mechanism due to the inclusion or exclusion of any individual's data [22]. In other words, ϵ governs the level of privacy protection provided by the differential privacy mechanisms, with lower values indicating stronger privacy guarantees.

Mathematically, a mechanism $M : X^n \rightarrow \mathbb{R}$ is ϵ-differentially private if $\forall$ datasets S, S' that differ on one entry, $\forall y \in \mathbb{R}$:

$$e^{-\epsilon} \leq \frac{Pr[M(S) = y]}{Pr[M(S') = y]} \leq e^{\epsilon} \tag{1}$$

[22]. This formulation bounds the probability of substantially different outputs for datasets differing in a single data point by a factor of e^{ϵ}. It ensures that the impact of any individual's data on the output is limited within a specific range determined by ϵ. As ϵ decreases, privacy protection strengthens, and when $\epsilon = 0$, the output is perfectly private, i.e. completely random noise. Conversely, increasing ϵ diminishes epsilon, and when $\epsilon \rightarrow \infty$, the output is perfectly accurate, i.e. no privacy protections at all.

In addition to ϵ-differential privacy, another concept within the field is approximate differential privacy, also known as (ϵ, δ)-differential privacy. This introduces an additional parameter δ that loosens the restrictions by allowing the mechanism to fail ϵ-differential privacy with probability δ [23]. Therefore, when $\delta = 0$, this is effectively the same as ϵ-differential privacy. To keep the discussion concise and focused, this paper will focus on ϵ-differential privacy, and all future mentions of "differential privacy" will be in reference to ϵ-differential privacy.

Having established the significance of ϵ in quantifying privacy guarantees, we briefly turn our attention to the mechanisms that translate ϵ into noise that can achieve these guarantees. Two widely used mechanisms in the realm of differential privacy are the Laplace and the Exponential mechanisms, which define different approaches of sampling noise according to a specified ϵ [25] [48].

In a machine learning context, prior to training the model, an overall privacy budget is determined, and then ϵ values are calculated by distributing the privacy budget over the number of passes through the data. During model training, privacy mechanisms inject ϵ-differentially private noise to the gradients. The resulting ML model consists of weights and biases that are differentially private and whose information leakage falls within the privacy budget [52].

Tradeoffs: Privacy vs Utility. When implementing differential privacy in the real-world, a crucial consideration is the tradeoff between privacy and utility. An increase in the level of privacy protection requires more noise to be added, which in turn results in a corresponding decrease in the accuracy and utility of the analysis [64].

A common approach to visualize this tradeoff is to plot privacy (inversely proportional to ϵ) against accuracy, as shown in Figure 4.

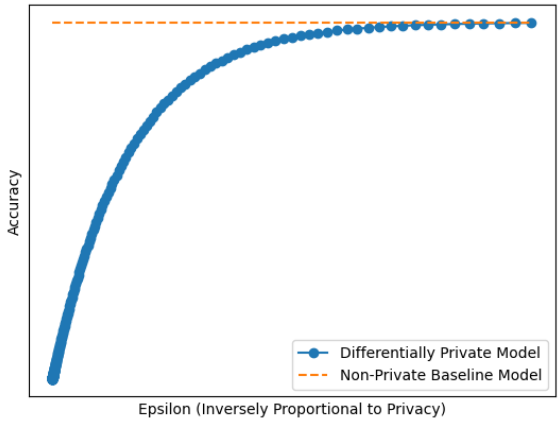


Fig. 4. Theoretical Privacy-Accuracy Tradeoff in Differential Privacy

In such graphs, as ϵ decreases (indicating stronger privacy guarantees), accuracy tends to decline, reflecting the increased noise introduced to maintain privacy. Conversely, as ϵ increases, accuracy improves at the expense of reduced privacy, plateauing as the accuracy approaches that of the non-private baseline model.

Different limitations and data contexts necessitate varying choices regrading the tradeoff privacy and accuracy. For instance, when handling sensitive patient data in healthcare, where privacy concerns are paramount, a higher level of privacy is essential to safeguard individual confidentiality. Conversely, in scenarios involving less sensitive information, such as age demographics, a lower level of privacy may suffice, allowing for greater accuracy in data analysis.

Moreover, different goals and tasks within an application further dictate the appropriate balance between privacy and accuracy. In applications focused on high-level statistical trends, a certain degree of accuracy may be sacrificed as long as general patterns are retained to maintain robust privacy protections. On the other hand, tasks requiring precise individual-level predictions demand higher accuracy levels, often at the expense of reduced privacy guarantees.

Real-world applications offer insights into how varying epsilon values are implemented to balance privacy and utility. For example, Apple's QuickType Suggestions and Emoji suggestions, which aim to provide personalized keyboard predictions while maintaining user privacy, operate using ϵ-DP with ϵ values of 8 and 4 respectively [1]. In scenarios where more stringent privacy guarantees are paramount, smaller ϵ values are employed. Google's Urban Mobility Dataset, which analyzes individual movement trends, utilizes (ϵ, δ)-DP with $\epsilon = 0.66$ [13], and the US Census Bureau's Post-Secondary Employment Outcomes Dataset leverages ϵ-DP with $\epsilon = 1.5$ [27]. For a more comprehensive list of such examples, refer to Table 8 in Appendix Section A.2.

2.4 Comparative Analysis of Existing Privacy-Preserving Approaches

Homomorphic encryption, federated learning, and differential privacy are three distinct approaches to preserving privacy in machine learning. Homomorphic encryption allows computation on encrypted data, preserving privacy but with computational overhead. Federated learning enables collaborative model training without sharing raw data, but requires communication and computation overhead. Differential privacy quantifies and guarantees privacy protections in data analysis, but may reduce utility and not protect against all attacks. Each approach offers unique benefits and

Approach	*What?*	*Why?*	*How?*	*At What Cost?*
Homomorphic Encryption	Allows computation on encrypted data	Protects sensitive data at rest and during computation	Utilizes cryptographic techniques and number theory	Computational overhead and reduced efficiency
Federated Learning	Enables collaborative model training without raw data sharing	Preserves privacy when aggregating model updates	Employs local model training and secure aggregation protocols	Computational overhead and reduced efficiency
Differential Privacy	Quantifies and guarantees privacy protections in data analysis	Ensures privacy of the training data in the model output	Injects randomness during model training to mask individual contributions	Reduced accuracy and utility

Table 2. Privacy-Preserving Approaches in the ML Pipeline

tradeoffs based on the specific requirements of privacy, computational resources, and utility, which are summarized in Table 2.

The tradeoffs between these approaches are best analyzed along the axes of privacy, accuracy, and efficiency, as shown in Figure 5. HE spans the privacy-efficiency plane, where the robustness of the chosen encryption scheme dictates the tradeoff between privacy and computational efficiency. FL leverages the concept of decentralized model training, offering privacy at the cost of efficiency. DP operates within the privacy-accuracy plane, where ϵ values govern the level of privacy vs utility.

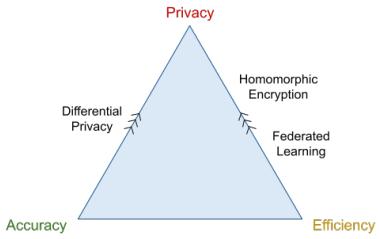
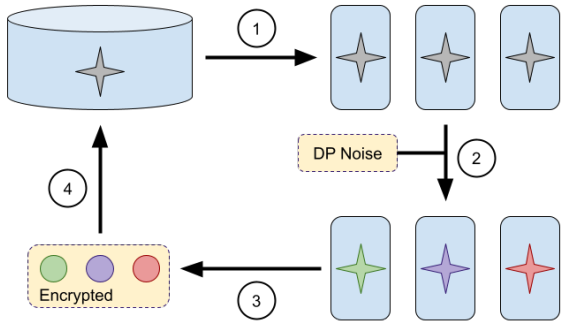


Fig. 5. Privacy-Accuracy-Efficiency Tradeoffs in Privacy-Preserving Machine Learning

Hybrid approaches that combine HE, FL, and DP offer enhanced privacy protections while addressing specific use case requirements. A privacy-centric strategy might leverage HE for secure data processing, FL for decentralizing model training, and DP for rigorous post-processing privacy guarantees. The process begins with a central machine learning model that is distributed to multiple devices participating in the training process. Each device independently trains its local model using its respective data. To ensure privacy during training, DP techniques are applied by adding controlled noise to the model updates. After training, the locally-generated and differentially-private model updates are aggregated and then encrypted using a HE scheme. Finally, the encrypted model updates are transmitted back to the central server, where they are applied to the central model. This integrated approach, outlined in Figure 6, ensures robust privacy protection and data confidentiality throughout the machine learning pipeline, from training to model update aggregation and application.

It is important to note that in hybrid designs involving both homomorphic encryption and federated learning, the encryption approach differs slightly from that of traditional homomorphic encryption in machine learning settings. Typically, in HE-ML scenarios, raw data is encrypted to protect privacy from the data curators. However, in federated

settings, this is not necessary, as the raw data never remains on local devices and is not transmitted. Instead, the focus shifts to encrypting the model updates, which are sent from devices to the central model. Encrypting these updates is crucial because they are susceptible to data reconstruction attacks or gradient inversion attacks, potentially compromising the confidentiality of the underlying data [36].



(1) Central server transmits central model to devices
(2) Devices train their models on local sensitive data in a DP manner
(3) Model updates are aggregated and encrypted according to a HE scheme
(4) Encrypted aggregated model updates are sent to the central server and applied to the central model

Fig. 6. Federated Differentially-Private ML with Homomorphic Encryption

2.5 Memorizing Models for Synthetic Data Generation

This section explores a novel approach to privacy preservation, initially introduced by Gerald Friedland and Jeffrey Bohn, which, despite its potential, has yet to be discussed in academic literature [29].

Motivation: Enhancing Privacy through Synthetic Data Generation. Synthetic data generation represents a promising strategy for preserving privacy, offering a solution to conventional risks associated with sensitive information by creating artificial datasets that mirror the statistical properties of real data without disclosing individuals' personally identifiable information (PII) [8].

The core principle of synthetic data generation lies in its capacity to produce new data points closely resembling the original dataset while upholding privacy standards. By substituting sensitive attributes with synthetic equivalents, data can be shared and analyzed without compromising individual privacy. This approach serves as a practical remedy for situations where access to authentic data is restricted due to privacy concerns, empowering researchers, businesses, and policymakers to derive insights and develop algorithms while respecting individuals' privacy rights.

Ultimately, the motivation for embracing synthetic data generation stems from its potential to fortify privacy safeguards in data-driven applications, bridging the gap between the demand for data-driven insights and the necessity to protect individual privacy.

Conceptual Understanding: Leveraging Memorizing Models for Synthetic Data. This novel approach leverages memorizing models for synthetic data generation, outlined both visually in Figure 7 and textually below:

(1) *Anonymization and Model Creation*: The data owner initiates the process by anonymizing the private data and constructing a memorizing model (MM). As part of the anonymization step, data is substituted with numbers

isomorphically, column-by-column, so that the per-column information content remains the same. A model is then trained on this data and is intentionally designed to overfit to it, capturing its intricate underlying patterns and relationships.

(2) *Synthetic Data Generation*: Leveraging the fundamental attributes of the private data, such as the minimum and maximum values of each feature, the data owner generates a synthetic dataset. These basic properties serve as guiding parameters in crafting synthetic counterparts that mirror the statistical characteristics of the original data. Notably, this step offers considerable flexibility tailored to the specific requirements of the use case. For instance, if minimum and maximum values are considered sensitive, mean and standard deviation could serve as alternate parameters. Depending on the chosen attributes, synthetic data can be generated using various methods, such as random sampling from uniform or normal distributions, or even through more intricate procedures tailored to specific dataset characteristics.

(3) *Labeling with Memorizing Model*: The memorizing model, having been trained on the original private data, is then employed to label the synthetic dataset. Drawing upon its learned insights, the model assigns labels to the synthetic instances.

(4) *Predictive Model Development*: Subsequently, the labeled synthetic data can be shared publicly with other data scientists, who can utilize it to construct predictive models. This step also offers significant flexibility. Because the data itself is being shared, there are no constraints imposed on the type of model that can data scientists can use.

(5) *Model Use and Evaluation*: The models trained on the labeled synthetic dataset can now be used for predictive tasks. Remarkably, these models demonstrate a notable similarity similarity, and in some instances, superior performance when compared to models trained using the original private data[2]. This underscores the potential of harnessing memorizing models for synthetic data generation in replicating the predictive capabilities of the original dataset while safeguarding individual privacy.

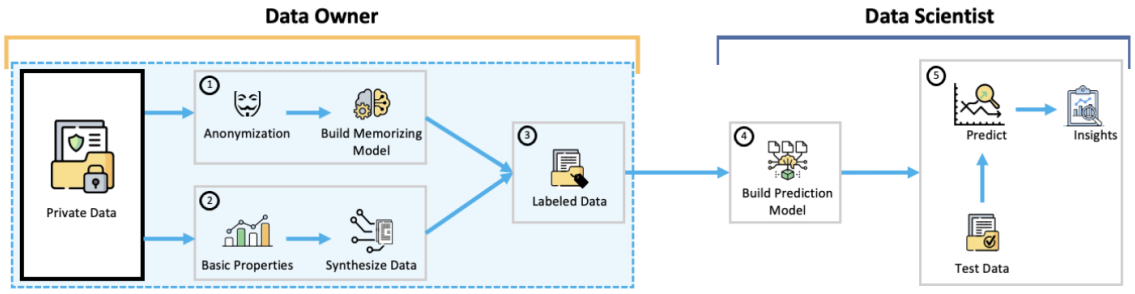


Fig. 7. Memorizing Models for Synthetic Data Generation Overview

Theoretical Underpinnings: Memorizing Models. In this section, we explore the theoretical basis of overfitting models in machine learning, with a specific focus on the application to the memorizing models for synthetic data generation approach.

[2]Comprehensive results can be explored in Section 3 regarding Case Studies

The concept of training a model to memorize the data while minimizing the number of parameters forms the cornerstone of this theoretical framework. This idea is fundamental to the concept of Memory Equivalent Capacity (MEC), which quantifies the maximum number of decisions or labels that a model can memorize from the training data. It serves as a comprehensive measure for evaluating the memorization capabilities of ML models, encompassing factors such as parameter count, model architecture, regularization techniques, and training imperfections.

Furthermore, the Generalization Ratio, introduced by Friedland, plays a crucial role in refining our understanding of model generalization within this context [28]. Defined as the ratio of correctly predicted bits to the memory equivalent capacity ($G = \frac{\#\ of\ correctly\ predicted\ bits}{memory\ equivalent\ capacity}$), this ratio serves as a quantitative metric for assessing the degree of overfitting exhibited by a model. Higher values of G signify better generalization and reduced overfitting, while lower values suggest a higher degree of memorization.

When training the memorizing model for synthetic data generation, the objective is to maximize the MEC to effectively capture the intricate patterns present in the original data without unnecessary complexity. The Generalization Ratio serves as a valuable metric in this context that helps quantify the associated privacy by giving insight into the degree of reversibility. The overall aim is to achieve a balance where the model effectively memorizes the data's labels while still exhibiting robust generalization, thereby safeguarding privacy without sacrificing utility.

Tradeoffs: Privacy vs Utility. This approach inherently entails a delicate balance between privacy and utility.

Central to this balance is the Generalization Ratio. Higher G values suggest stronger generalization capabilities, enhancing privacy but potentially compromising accuracy by overlooking nuanced data patterns.

The approach seeks to strike this balance by optimizing the memorization of data labels while minimizing model complexity. By leveraging the MEC to guide the memorization process, the goal is to preserve privacy without sacrificing the predictive power of the model.

However it is important to acknowledge that further research is necessary t quantify the extent of privacy guarantees empirically. This entails exploring measurable privacy metrics and assessing how different methods of synthetic data generation can impact the level of privacy preservation.

Overall, utilizing MMs for synthetic data generation requires a nuanced understanding of the privacy-preserving technique and its implications for model utility and privacy protection.

3 CASE STUDIES

In this section, we delve into real-world case studies aimed at dissecting the performance and identifying the pain points associated with implementing various privacy-preserving techniques in machine learning applications. Through these case studies, we explore how federated learning, differential privacy, the combination of the two, as well as the novel approach of memorizing models for synthetic data generation fare in practical scenarios, shedding light on their effectiveness, limitations, and areas for improvement[3]. By examining these case studies, we aim to provide valuable insights into the viability and applicability of these techniques in addressing privacy concerns in machine learning models.

3.1 Datasets

We used a variety of datasets in the case studies to illustrate the applications of federated learning, differential privacy, and memorizing models for synthetic data generation, which are summarized in Table 3. These datasets have been

[3]Homomorphic encryption has already been evaluated extensively in existing research literature

carefully selected to showcase various use cases and demonstrate both the effectiveness and limitations of different privacy-enhancing methodologies in real-world scenarios.

3.1.1 UCI Student Performance [17]. This dataset contains data on student grades, demographics, and social and school-related features for secondary school students from two Portuguese schools in 2008. With 30 features covering 649 students, the regression task aims to predict a student's final-period math grades. We use 435 (67%) randomly chosen students to train the model. Given that this dataset includes sensitive information such as student grades, demographics, and social attributes, ensuring privacy is essential to safeguard the confidentiality and integrity of individuals' personal data.

3.1.2 King County House Sales [4]. This Kaggle dataset contains sale prices for houses in King County, Washington sold between May 2014 and May 2015, alongside 20 additional features describing various quantitative and qualitative characteristics of each property. With data spanning over 21,000 houses, the regression task aims to predict the value of a given home in the area. For model training, 14,480 randomly selected observations, representing 67% of the total dataset are utilized. Given the sensitive nature of property prices and related attributes, ensuring privacy during model training is crucial to protect the confidentiality of homeowner's information.

3.1.3 UCI Adult (Census Income) [14]. This dataset comprises census data from the 1994 US Census, encompassing 14 features that represent demographic and income information. 32,724 instances, constituting 67% of the original 48,000+, are utilized to train a classification model to predict whether a given adult would have an annual income exceeding $50,000. Maintaining model privacy in this context is paramount to safeguard the confidentiality of individuals' income and demographic details.

3.1.4 UCI Individual Household Electric Power Consumption [32]. This dataset comprises over 2 million total measurements collected from a household in France between December 2006 and November 2010, representing electric power consumption at a one-minute sampling rate. There are 9 features, representing different electrical quantities and some sub-metering values. The regression task involves predicting the Global Active Power of the house using the available features. Privacy is important in this scenario to ensure the confidentiality of household energy consumption patterns.

3.1.5 UCI Wisconsin Breast Cancer [63]. This dataset contains information regarding the characteristics of cell nuclei present in breast masses of over 550 breast cancer patients in Wisconsin. It includes 30 features, ranging from radius and perimeter measurements to smoothness and compactness. The classification task entails predicting whether these tumors are malignant or benign. Privacy is critical in this context, due to the sensitive nature of the data. It contains detailed medical information of patients that could be misused if not handled securely.

3.1.6 UCI Handwritten Digits [11]. This dataset features a total of 5620 handwritten digits between 0 and 9, collected by the National Institute of Standards and Technology (NIST). The original 32x32 images were pre-processed into a set of 64 features, each ranging from 0 to 16. These features represent the pixel count within individual, non-overlapping 8x8 blocks. The classification task is to identify the digit written in each image. Although privacy is not a primary concern with this dataset, it serves as a valuable benchmark for multi-class classification tasks.

3.2 Methodology

3.2.1 Federated Learning.

Dataset	*Description*	*Dimensions*	*Task*	*Approach Tested*
UCI Student Performance	Predict student grades given demographic and academic features	(649, 30)	Regression	Differential Privacy
King County House Sales	Predict house prices for properties given quantitative and qualitative features	(21613, 21)	Regression	Differential Privacy
UCI Adult (Census Income)	Predict whether an individual would have an annual income exceeding $50k given demographic features	(48842, 14)	Classification (Binary)	Differential Privacy and MMs for Synthetic Data Generation
UCI Household Power Consumption	Predict global active power of a household given energy consumption features	(2075259, 9)	Regression	Federated Learning
UCI Wisconsin Breast Cancer	Predict whether breast cancer tumors are malignant or benign given the characteristics of the cell nuclei	(569, 30)	Classification (Binary)	MMs for Synthetic Data Generation
UCI Handwritten Digits	Predict handwritten digits (0-9) from pixel information	(5620, 64)	Classification (Multi)	MMs for Synthetic Data Generation

Table 3. Datasets used in Case Studies

(1) *Pre-Processing*: To prepare the dataset for analysis, standard pre-processing techniques were applied in order to enhance data quality and optimize model performance. This included handling missing values, scaling numerical features, and separating the data into training and testing sets.
(2) *Evaluating Performance*: Federated learning was simulated using mini-batch Stochastic Gradient Descent (SGD) with a batch size of 1 to mimic individual devices learning autonomously. Additionally, a random delay was introduced to simulate the transmission time for relaying model updates back to the central model. This approach emulated the decentralized training environment typical of federated learning scenarios. The federated model was then compared against a baseline model trained using mini-batch SGD with a batch size of 32 and no delay. The comparison was conducted across various dataset sizes (number of rows) to evaluate how the performance of the federated learning model scaled with increasing data volume.

3.2.2 *Differential Privacy.*

(1) *Pre-Processing*: The first step involved pre-processing the dataset to prepare it for analysis. This included tasks such as one-hot encoding categorical variables to convert them into numerical representations, splitting the dataset into training and testing sets, and scaling the dataset to ensure uniformity and stability in model training.
(2) *Evaluating Dimensionality*: The baseline model was compared to the private model at various levels of dimensionality to assess the impact of privacy-preserving techniques on model performance. Private models were implemented using IBM's DiffPrivLib Python Library, which provides various differential privacy tools and mechanisms, as well as DP implementations of several types of machine learning models [34]. Both models were evaluated using varying numbers of features to analyze how the introduction of privacy affected model performance across different dimensionalities. The k-most influential features were selected by picking those with the largest associated absolute coefficients from a basic linear/logistic regression model. It is important to note that in practical scenarios, when fitting a basic model may not be appropriate, there are private methods available for feature selection, such as private PCA [35].

(3) *Evaluating ϵ (Privacy Parameter)*: For a fixed set of features, the performance of both the baseline model and the private model was evaluated across different values of ϵ. By varying ϵ, the impact of privacy constraints on model performance and decision-making outcomes could be assessed. This analysis provided insights into the trade-offs between privacy and utility, helping to inform the selection of an appropriate ϵ value for the practical deployment of the private model.

3.2.3 Differentially-Private Federated Learning (FL+DP). This hybrid approach was modeled using differentially-private mini-batch Stochastic Gradient Descent [7]. The methodology mirrors the steps outlined previously, encompassing dataset pre-processing, evaluation of model training time efficiency, evaluation of model performance across varying dimensions, and evaluation of model performance across different privacy parameters.

For this implementation of DP mini-batch SGD, it is important to note that instead of ϵ, we examined two crucial parameters: noise and bound. Noise denotes the level of noise introduced to the gradient at each iteration, while bound represents the graident norm bound, determining the maximum allowable update to the gradient at any given step. While noise has a direct mathematical mapping to ϵ, it is intriguing to note that the bound also influences model performance in this approach, which is why we analyzed various combinations of the two.

3.2.4 MMs for Synthetic Data Generation. To assess this approach, we utilize the Brainome package [5], which offers built-in support for generating memorizing models and simultaneously trains various types of models (RF, NN, DT, SVM) to allow for ease of comparison.

The process starts by preprocessing the dataset, one-hot encoding categorical variables to transform them into numerical representations. Subsequently, ML models were trained on this preprocessed data to establish baseline metrics for comparison. Following this, we trained a decision tree-based memorizing model designed to overfit to the data.

We then generated 200,000 synthetic data points by sampling from a uniform distribution using the minimum and maximum values of each features. These synthetic data points were labeled using the previously trained MM. Finally, ML models were trained on the labeled synthetic data to evaluate the approach's efficacy.

3.3 Evaluation

3.3.1 Federated Learning. In the evaluation of federated learning models, total execution time was the key metric. Total execution time encompasses the duration required for both the training phase on individual devices and the simulated communication delays for relaying model updates. This metric is critical in assessing the efficiency and scalability of federated learning systems, especially as the dataset size and number of participating devices increase.

3.3.2 Differential Privacy. In comparing the differentially private ML model to the baseline model, we employed a comprehensive set of metrics to evaluate their respective performance and effectiveness.

- *Standard Error Metric*: This metric provides a quantitative measure of the predictive accuracy of the private model compared to the baseline model, by looking at the standard error metric used for the predictive task. For regression tasks, Root Mean Squared Error (RMSE) was utilized to assess the deviation of predicted values from actual values. For classification tasks, accuracy was employed to measure the proportion of correctly classified instances. By comparing these metrics between the private and baseline models, we gain insights into the relative predictive performance of each approach.
- *Difference Between Model Parameters*: This metric allows us to assess the degree of divergence between the parameters learned by the private model and those of the baseline model. This only applies to parameterized

models like linear/logistic regression, as the coefficients of the same features between models can be directly compared to one another. By analyzing these differences, we can identify how privacy-preservation influences the learned model parameters, and thereby the accuracy of the models.

- *Results of Hypothetical Decisions*: In addition to evaluating the predictive accuracy and model parameters, we also examined the practical implications of model predictions on decision-making processes [38, 53]. We simulated hypothetical decision-making scenarios, such as allocating subsidies or offering tutoring based on model predictions (from the Income and Student Performance datasets respectively), and compared the outcomes between the private and baseline models. This qualitative assessment offers a deeper understanding of how privacy-preserving mechanisms may impact real-world decision-making processes and outcomes.

3.3.3 Differentially-Private Federated Learning (FL+DP). In assessing the FL+DP models, our objective was to analyze both time efficiency and model performance across different dimensions, observations, and model parameters. We independently evaluated time efficiency across observations and model performance across dimensionality. Subsequently, we examined the time efficiency and model performance across observations for different models trained using various noise and bound parameter combinations.

3.3.4 MMs for Synthetic Data Generation. Since the Brainome package currently only supports classification tasks, we evaluate metrics that are specific to classification tasks. Specifically, we analyze the validation accuracy of the top-performing ML model trained on the raw data vs the top-performing ML model trained on the synthetic data. We also analyze the Generalization Ratios for the memorizing models trained on the original data, as it gives insights into the level of privacy the approach provides.

3.4 Results

3.4.1 Federated Learning. In the simulation of federated learning, a random delay was incorporated by sampling from a normal distribution with a mean of 1 ms and a standard deviation of 0.5 ms (non-negative). The results, including time efficiency and performance metrics across varying numbers of observations, are shown in Table 4.

As expected, federated learning exhibits increased execution time as the number of observations (simulated devices) increases. This correlation is due to our modeling approach, where each observation represents a device and the simulated delay per device accumulates with more observations.

A notable observation from the results is that the performance of the federated model remains relatively stable with increasing numbers of observations, whereas the baseline model shows an initial poor performance that improves with larger datasets.

The difference in performance trends can be attributed to the differences in training methodologies. The baseline model's batch size of 32 allows it to process more data per iteration, potentially leading to faster convergence but more fluctuation on smaller datasets. In contrast, the federated model's batch size of 1 necessitates more iterations for training but results in a more consistent performance across varying dataset sizes.

3.4.2 Differential Privacy. Differential privacy experiments were conducted across various datasets and multiple machine-learning models. This section will focus on the performance of differentially-private linear regression models trained on the King County House Sales dataset as an instructive example, but the full results of the experiments can be found in Appendix Section A.3.

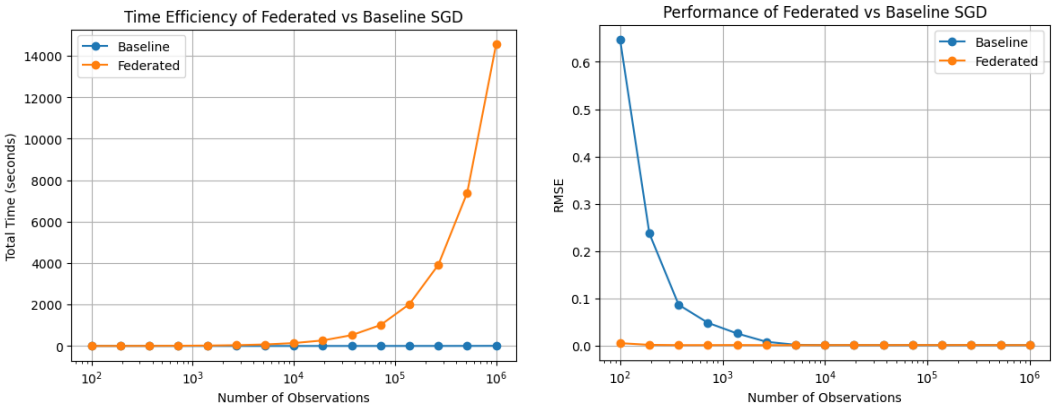


Table 4. Federated vs Baseline Stochastic Gradient Descent (Household Power Dataset)

The specific metrics analyzed include (1) the root mean squared error (RMSE) of home value predictions, (2) the distance between the private model's coefficients and the baseline model's coefficients, and (3) the difference between the private model's forecasted property tax and the true property tax (simulated as a simple function of home value).

Table 5 highlights these metrics across varying numbers of features. Notably, the results are definitive, yet counterintuitive: as the number of features increases, the model consistently performs worse across all three metrics. Despite the dataset having 141 features after pre-processing, when zooming in on the graph, it seems that the model actually performs best with just 1 feature.

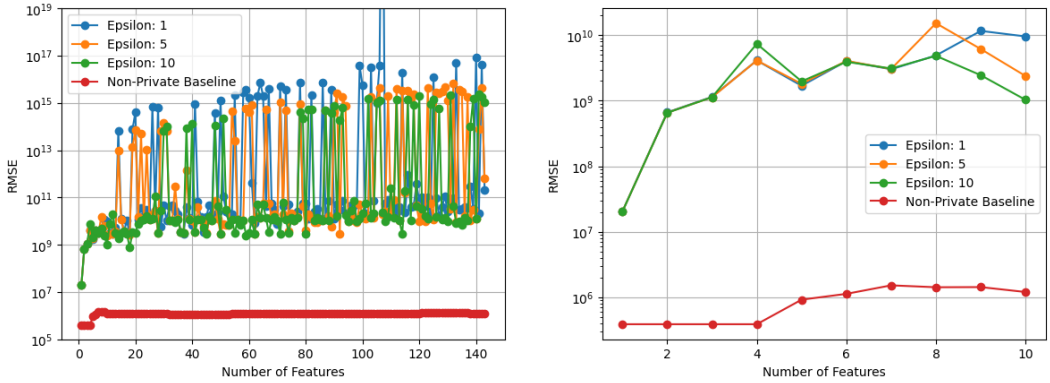
To delve deeper into this phenomenon, three new models were trained on datasets consisting of the top 1, 3, and 10 features, shown in Table 6. These results reaffirm our earlier findings, as the model trained on data with a single feature outperforms those with 3 and 10 features.

This observed trend of model performance decreasing with higher dimensionality persists across different datasets, machine learning tasks, and model types. Often, the highest-performing models leverage a seemingly unreasonably small number of features for a complex task. This challenges traditional machine learning paradigms, where more features generally translate to improved model performance and accuracy.

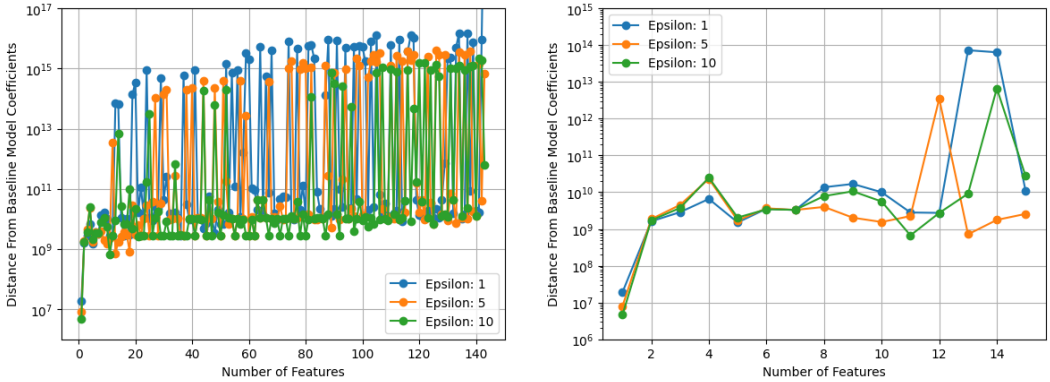
Here lies one of the primary takeaways of this paper: privacy-preserving machine learning exhibits a "Curse of Dimensionality."

To better understand this, we go back to the definition of differential privacy in the context of machine learning, where it denotes the inability to confidently determine whether a specific individual was included in the training dataset. As the dataset dimensionality increases, preserving individual privacy becomes increasingly challenging. With more features, more information is collected that can be used to uniquely identify individuals. Consequently, to maintain differential privacy, more noise must be injected to try to mask the identification of individuals. While this noise helps to ensure privacy, high levels can compromise the utility of the trained models, thereby obscuring valuable insights behind the magnitude of the noise. Overall, these experiments underscore the delicate balance between privacy preservation and model utility in machine-learning contexts, as well as the nuanced considerations involved in real-world implementation.

RMSE of DP vs Baseline Linear Regression



Distance of DP Coefficients from Baseline Linear Regression



Simulated Property Tax Difference of DP vs Baseline Linear Regression

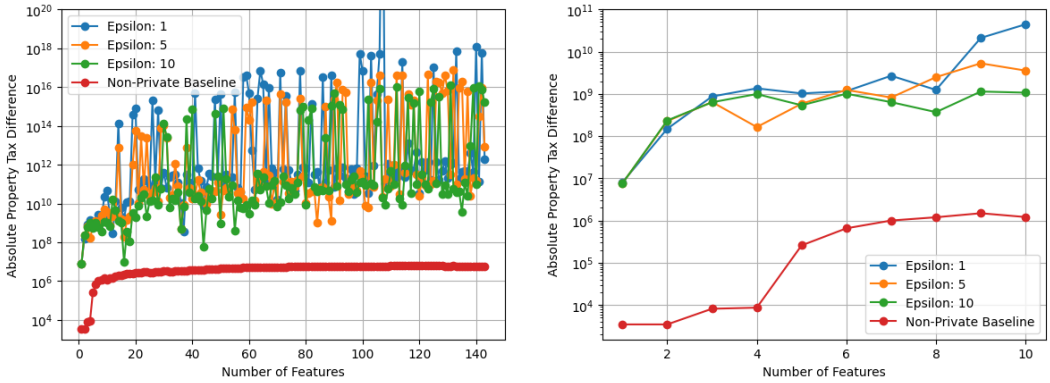


Table 5. DP vs Baseline Linear Regression Model for Different Feature Set Sizes (House Prices Dataset)

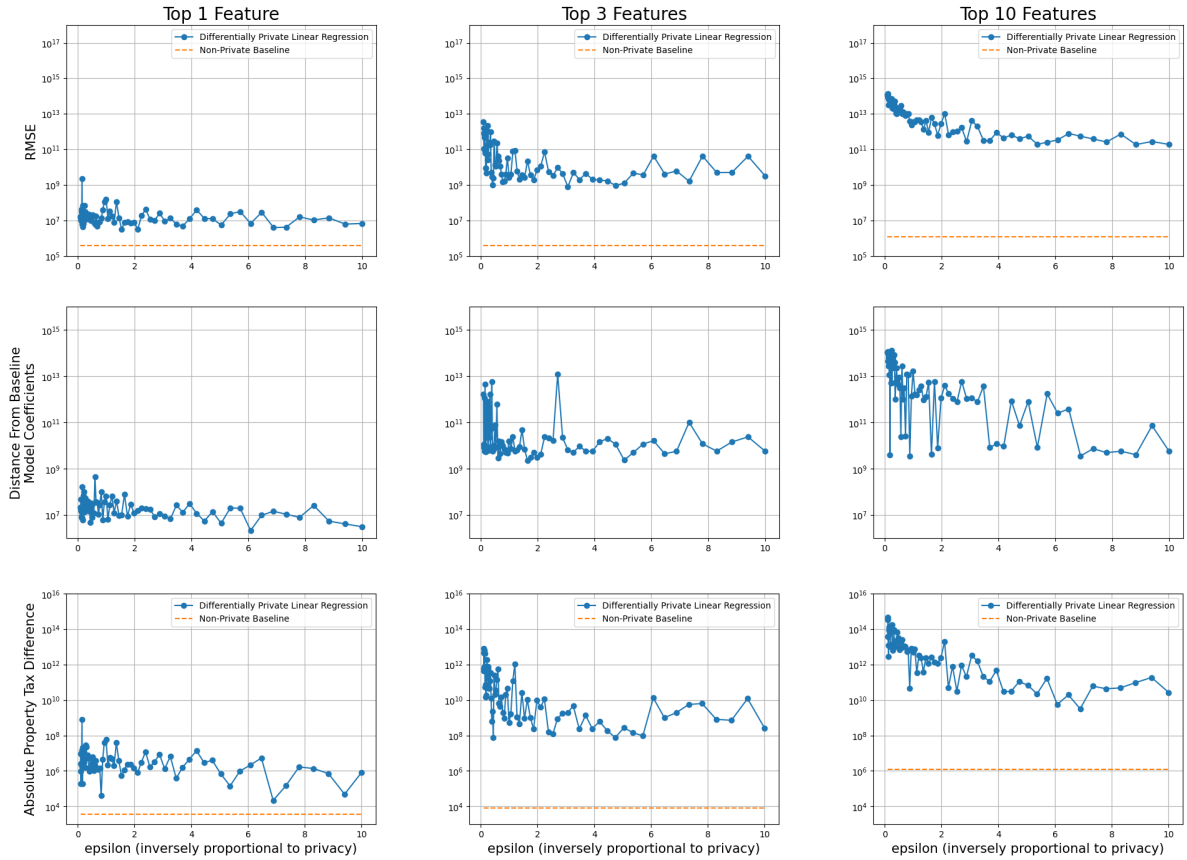


Table 6. DP vs Baseline Linear Regression Model using Top 1, 3, 10 Features (House Prices Dataset)

3.4.3 Differentially-Private Federated Learning (FL+DP). In evaluating the FL+DP models, we simulated the federated aspect of model training in the same manner, by incorporating a random delay sampled from a normal distribution with a mean of 1ms and a standard deviation of 0.5ms (non-negative).

The key highlight of our results pertains to the performance of different models using various combinations of noise and bound parameters across observations. These results are summarized in Figure 8, where the y-axis represents the RMSE, the x-axis represents the number of observations, and the baseline and FL+DP models are denoted in yellow and blue respectively.

Intriguingly, the FL+DP model demonstrates robust performance across most scenarios, with the exception of the bottom right quadrant of the table, indicating the intersection of high noise and high bound parameters. This is because when high noise is coupled with low bound, the impact of noise on gradients is limited, thus mitigating its adverse effects on model performance. Conversely, when noise is high and the bound is also high, the noise can significantly alter the gradients, leading to notable poor model performance.

The remaining results align closely with expected behavior and are detailed in Appendix Section A.3.

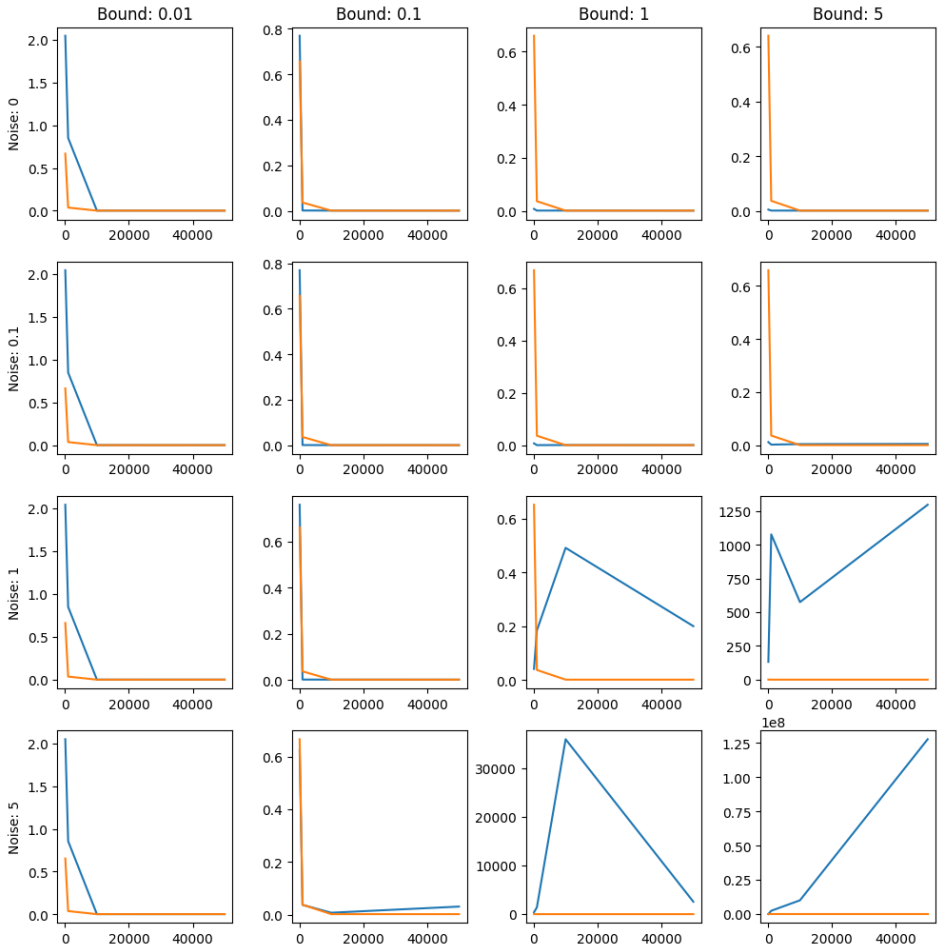


Fig. 8. RMSE of FL-DP vs Baseline Model (Household Power Dataset)

3.4.4 MMs for Synthetic Data Generation. Intriguingly, this approach consistently resulted in considerable improvements in validation accuracy across all evaluated datasets, spanning both binary and multi-class classification tasks. The results are illustrated in Table 7.

The Generalization Ratios for the models trained on the synthetic data also demonstrate promising levels of privacy preservation. Particularly striking are the high G values observed in the Wisconsin Breast Cancer dataset and the Handwritten Digit dataset, reaching 7.42 and 12.36 bits/bit respectively. Despite the elevated privacy measures, these models still achieved near-100% accuracy.

One possible explanation for this remarkable performance could be attributed to the inherent flexibility of the approach. By generating synthetic data points, potentially infinitely many, nuanced patterns and trends from the

Dataset	*MM Generalization Ratio*	*Original Model Validation Accuracy*	*Synthetic Model Validation Accuracy*	*Improvement*
UCI Adult (Census Income)	2.42 bits/bit	87.34% (RF)	97.04% (DT)	+9.70%
UCI Wisconsin Breast Cancer	7.42 bits/bit	96.49% (RF)	99.99% (DT)	+3.5%
UCI Handwritten Digits	12.36 bits/bit	96.54% (RF)	99.98% (DT)	+3.44%

Table 7. Results of Baseline Approach vs Using MMs for Synthetic Data Generation

original dataset are amplified. This expanded dataset may enable the model to grasp the complex relationships more effectively, leading to superior predictive capabilities.

These findings collectively underscore the efficacy of this approach not only in enhancing model performance, but also in bolstering privacy protection, thus offering a promising solution for privacy-sensitive applications.

4 DISCUSSION

The results of our experiments shed light on the intricate dynamics of privacy-preserving machine learning in practice, highlighting both challenges and opportunities in real-world implementation. In federated learning, we observed that while the execution time increased with the number of observations (simulated devices), the federated model exhibited stable performance across varying dataset sizes. This underscores the potential of federated learning for privacy-preserving model training, especially in scenarios where data privacy is paramount. However, the tradeoff between execution time and model performance warrants further investigation, particularly in optimizing communication protocols and modeling simultaneous training on distributed devices to enhance efficiency without compromising accuracy.

On the other hand, our experiments on differential privacy revealed a noteworthy phenomenon: the "Curse of Dimensionality" in privacy-preserving model training. Despite conventional wisdom dictating that more features lead to better model performance, we found that increasing dimensionality consistently resulted in poorer model performance across various metrics. This underscores the delicate balance between privacy preservation and model utility, as injecting too much noise to ensure differential privacy can obscure valuable insights.

Additionally, the remarkable improvements in validation accuracy and promising levels of privacy preservation achieved through the Memorizing Models for Synthetic Data Generation approach carry significant implications for PPML implementations. By demonstrating enhanced model performance without compromising individual privacy, this approach offers a practical solution for organizations and practitioners handling sensitive data. The high Generalization Ratios underscore the balance between privacy preservation and model utility, empowering analysts to derive accurate predictions while upholding privacy standards.

These findings have significant implications for the practical implementation of PPML techniques. While federated learning shows promise for decentralized model training without compromising data privacy, careful optimization is necessary to minimize execution time. Similarly, differential privacy presents challenges in maintaining model utility while ensuring individual privacy, highlighting the need for innovative approaches to mitigate the "Curse of Dimensionality" and preserve valuable insights in high-dimensional datasets. Finally, further research is necessary to better understand the limitations of utilizing MMs for Synthetic Data Generation.

5 CONCLUSION

This thesis has delved into the practical intricacies of privacy-preserving machine learning, examining the effectiveness of various techniques in real-world scenarios. Through experiments on federated learning and differential privacy, we have uncovered key insights into the tradeoffs between privacy, utility, and efficiency in PPML. Our findings underscore the importance of balancing these factors to develop robust and privacy-compliant machine learning models.

Looking ahead, further research is needed to address the challenges identified in our experiments and advance the field of PPML. Future work may focus on optimizing federated learning protocols for improved efficiency and scalability, as well as developing novel techniques to mitigate the "Curse of Dimensionality" in differential privacy. Further research is essential to investigate the limitations, scalability, and privacy guarantees of the MMs for Synthetic Data Generation approach. Additionally, exploring hybrid approaches that combine homomorphic encryption, federated learning, differential privacy, MMs for synthetic data generation and/or other privacy-preserving techniques could offer enhanced privacy protections while preserving model utility in complex real-world applications.

Overall, this thesis contributes to the growing body of knowledge in PPML, providing valuable insights and paving the way for future research and innovation in this critical area. By prioritizing privacy throughout the machine learning lifecycle, we can empower practitioners to develop secure and privacy-compliant models that safeguard sensitive data while enabling valuable insights and applications.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my primary thesis advisor, Nitin Kohli, for his unwavering support, invaluable guidance, and insightful feedback throughout the entire research process. This thesis would not have been possible without his help.

I am also grateful to my thesis advisor, Professor Gerald Friedland, for his valuable input and feedback, especially for his contributions pertaining to the Memorizing Models for Synthetic Data Generation approach. His knowledge and expertise have been instrumental in shaping this work.

Additionally, this endeavor would not have been possible without the generous support from Professors Eric Van Dusen and Narges Norouzi. I would also like to acknowledge and thank my peers in DATA H195A/B for their ongoing feedback and support, which have been instrumental in shaping my research.

REFERENCES

[1] [n. d.]. https://www.apple.com/privacy/docs/Differential_Privacy_Overview.pdf
[2] 2014. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
[3] 2016. https://gdpr-info.eu/
[4] 2016. https://www.kaggle.com/datasets/harlfoxem/housesalesprediction
[5] 2022. https://www.brainome.ai/
[6] 2023. https://www.ctdatacollaborative.org/global-victim-perpetrator-synthetic-dataset
[7] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.
[8] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. 2019. Privacy preserving synthetic data release using deep learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18*. Springer, 510–526.
[9] Abbas Acar, Hidayet Aksu, A Selcuk Uluagac, and Mauro Conti. 2018. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Computing Surveys (Csur)* 51, 4 (2018), 1–35.
[10] Fernando Almeida. 2018. Big data: concept, potentialities and vulnerabilities. *Emerging Science Journal* 2, 1 (2018), 1–10.
[11] E. Alpaydin and C. Kaynak. 1998. Optical Recognition of Handwritten Digits. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C50P49.

[12] Ho Bae, Jaehee Jang, Dahuin Jung, Hyemi Jang, Heonseok Ha, Hyungyu Lee, and Sungroh Yoon. 2018. Security and privacy issues in deep learning. *arXiv preprint arXiv:1807.11655* (2018).

[13] Aleix Bassolas, Hugo Barbosa-Filho, Brian Dickinson, Xerxes Dotiwalla, Paul Eastham, Riccardo Gallotti, Gourab Ghoshal, Bryant Gipson, Surendra A Hazarie, Henry Kautz, et al. 2019. Hierarchical organization of urban mobility and its connection with city livability. *Nature communications* 10, 1 (2019), 4817.

[14] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[15] Eleanor Birrell, Jay Rodolitz, Angel Ding, Jenna Lee, Emily McReynolds, Jevan Hutson, and Ada Lerner. 2023. SoK: Technical Implementation and Human Impact of Internet Privacy Regulations. *arXiv preprint arXiv:2312.15383* (2023).

[16] Joshua E Blumenstock and Nitin Kohli. 2023. Big Data Privacy in Emerging Market Fintech and Financial Services: A Research Agenda. *arXiv preprint arXiv:2310.04970* (2023).

[17] Paulo Cortez. 2014. Student Performance. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5TG7T.

[18] Giuseppe D'Acquisto, Josep Domingo-Ferrer, Panayiotis Kikiras, Vicenç Torra, Yves-Alexandre de Montjoye, and Athena Bourka. 2015. Privacy by design in big data: an overview of privacy enhancing technologies in the era of big data analytics. *arXiv preprint arXiv:1512.06000* (2015).

[19] Damien Desfontaines. 2023. A list of real-world uses of differential privacy. https://desfontain.es/privacy/real-world-differential-privacy.html

[20] Irit Dinur and Kobbi Nissim. 2003. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 202–210.

[21] Eric Durnell, Karynna Okabe-Miyamoto, Ryan T Howell, and Martin Zizi. 2020. Online privacy breaches, offline consequences: Construction and validation of the concerns with the protection of informational privacy scale. *International Journal of Human–Computer Interaction* 36, 19 (2020), 1834–1848.

[22] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12.

[23] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*. Springer, 486–503.

[24] Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. 2019. Differential privacy in practice: Expose your epsilons! *Journal of Privacy and Confidentiality* 9, 2 (2019).

[25] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284.

[26] Haokun Fang and Quan Qian. 2021. Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet* 13, 4 (2021), 94.

[27] Andrew David Foote, Ashwin Machanavajjhala, and Kevin McKinney. 2019. Releasing earnings distributions using differential privacy: Disclosure avoidance system for post-secondary employment outcomes (pseo). *Journal of Privacy and Confidentiality* 9, 2 (2019).

[28] Gerald Friedland. 2024. *Information-Driven Machine Learning Data Science as an engineering discipline.* Springer International Publishing AG.

[29] Gerald Friedland and Jeffrey Bohn. [n. d.]. Model Privacy: Sharing Data with an Enforced Purpose.

[30] Simson Garfinkel, John M Abowd, and Christian Martindale. 2019. Understanding database reconstruction attacks on public data. *Commun. ACM* 62, 3 (2019), 46–53.

[31] Sofie Goethals, Kenneth Sörensen, and David Martens. 2023. The privacy issue of counterfactual explanations: explanation linkage attacks. *ACM Transactions on Intelligent Systems and Technology* 14, 5 (2023), 1–24.

[32] Georges Hebrail and Alice Berard. 2012. Individual household electric power consumption. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C58K54.

[33] Amaç Herdağdelen, Alex Dow, Bogdan State, Payman Mohassel, and Alex Pompe. 2020. Protecting privacy in facebook mobility data during the COVID-19 response - meta research. https://research.facebook.com/blog/2020/06/protecting-privacy-in-facebook-mobility-data-during-the-covid-19-response/

[34] Naoise Holohan, Stefano Braghin, Pól Mac Aonghusa, and Killian Levacher. 2019. Diffprivlib: the IBM differential privacy library. *ArXiv e-prints* 1907.02444 [cs.CR] (July 2019).

[35] Hafiz Imtiaz and Anand D Sarwate. 2016. Symmetric matrix perturbation for differentially-private principal component analysis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2339–2343.

[36] Weizhao Jin, Yuhang Yao, Shanshan Han, Carlee Joe-Wong, Srivatsan Ravi, Salman Avestimehr, and Chaoyang He. 2023. FedML-HE: An Efficient Homomorphic-Encryption-Based Privacy-Preserving Federated Learning System. *arXiv preprint arXiv:2303.10837* (2023).

[37] Georgios Kaissis, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima Jr, Jason Mancuso, Friederike Jungmann, Marc-Matthias Steinborn, et al. 2021. End-to-end privacy preserving deep learning on multi-institutional medical imaging. *Nature Machine Intelligence* 3, 6 (2021), 473–484.

[38] Nitin Kohli, Emily Aiken, and Joshua Blumenstock. 2023. Privacy Guarantees for Personal Mobility Data in Humanitarian Response. *arXiv preprint arXiv:2306.09471* (2023).

[39] Satyam Kumar, Dayima Musharaf, Seerat Musharaf, and Anil Kumar Sagar. 2023. A Comprehensive Review of the Latest Advancements in Large Generative AI Models. In *International Conference on Advanced Communication and Intelligent Systems*. Springer, 90–103.

[40] Joseph Kupfer. 1987. Privacy, autonomy, and self-concept. *American Philosophical Quarterly* 24, 1 (1987), 81–89.

[41] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. 2021. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[42] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. 2021. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–36.

[43] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. 2018. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889* (2018).

[44] Adrienn Lukács. 2016. What is privacy? The history and definition of privacy. (2016).

[45] Jeff Luszcz. 2018. Apache struts 2: how technical and development gaps caused the equifax breach. *Network Security* 2018, 1 (2018), 5–8.

[46] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[47] Brendan McMahan and Abhradeep Thakurta. 2022. Federated learning with formal differential privacy guarantees. https://blog.research.google/2022/02/federated-learning-with-formal.html

[48] Frank McSherry and Kunal Talwar. 2007. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. IEEE, 94–103.

[49] Solomon Messing, Christina DeGregorio, Bennett Hillenbrand, Gary King, Saurav Mahanti, Zagreb Mukerjee, Chaya Nayak, and Nate Persily. [n. d.]. State, Bogdan; Wilkins, Arjun, 2020," Facebook Privacy-Protected Full URLs Data Set".

[50] Deirdre K Mulligan, Colin Koopman, and Nick Doty. 2016. Privacy is an essentially contested concept: a multi-dimensional analytic for mapping privacy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374, 2083 (2016), 20160118.

[51] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 111–125.

[52] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. 2023. How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research* 77 (2023), 1113–1201.

[53] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. 2020. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 189–199.

[54] Maria Rigaki and Sebastian Garcia. 2023. A survey of privacy attacks in machine learning. *Comput. Surveys* 56, 4 (2023), 1–34.

[55] Ryan Rogers, Subbu Subramaniam, Sean Peng, David Durfee, Seunghyun Lee, Santosh Kumar Kancha, Shraddha Sahay, and Parvez Ahammad. 2020. LinkedIn's Audience Engagements API: A privacy preserving data analytics system at scale. *arXiv preprint arXiv:2002.05839* (2020).

[56] Ahmed Salem, Apratim Bhattacharya, Michael Backes, Mario Fritz, and Yang Zhang. 2020. {Updates-Leak}: Data set inference and reconstruction attacks in online learning. In *29th USENIX security symposium (USENIX Security 20)*. 1291–1308.

[57] Sarah Hartman-Caverly and Alexandria Edyn Chisholm. 2023. From Data Harm to Data Justice: Privacy and Social Justice. (2023). https://doi.org/10.26207/QAWG-NK07

[58] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE, 3–18.

[59] Daniel Solove. 2014. 10 reasons why privacy matters. https://teachprivacy.com/10-reasons-privacy-matters/

[60] Daniel J Solove. 2005. A taxonomy of privacy. *U. Pa. l. Rev.* 154 (2005), 477.

[61] Spencer Wheatley, Annette Hofmann, and Didier Sornette. 2019. Data breaches in the catastrophe framework & beyond. *arXiv preprint arXiv:1901.00699* (2019).

[62] Spencer Wheatley, Thomas Maillart, and Didier Sornette. 2016. The extreme risk of personal data breaches and the erosion of privacy. *The European Physical Journal B* 89 (2016), 1–12.

[63] William Wolberg, Olvi Mangasarian, Nick Street, and W. Street. 1995. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5DW2B.

[64] Alexandra Wood, Micah Altman, Aaron Bembenek, Mark Bun, Marco Gaboardi, James Honaker, Kobbi Nissim, David R O'Brien, Thomas Steinke, and Salil Vadhan. 2018. Differential privacy: A primer for a non-technical audience. *Vand. J. Ent. & Tech. L.* 21 (2018), 209.

[65] Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. 2018. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903* (2018).

[66] Yin Yang, Zhenjie Zhang, Gerome Miklau, Marianne Winslett, and Xiaokui Xiao. 2012. Differential privacy in data publication and analysis. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. 601–606.

[67] Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. 2021. A survey on federated learning. *Knowledge-Based Systems* 216 (2021), 106775.

[68] Xiaojin Zhang, Yan Kang, Kai Chen, Lixin Fan, and Qiang Yang. 2023. Trading Off Privacy, Utility, and Efficiency in Federated Learning. *ACM Transactions on Intelligent Systems and Technology* 14, 6 (2023), 1–32.

A APPENDIX

A.1 GitHub Repository

The associated repository for this thesis serves as a comprehensive resource containing all the datasets utilized in the research, Python notebooks housing the code utilized for analysis and experimentation, and visualizations depicting the findings. Accessible through https://github.com/AneeshPatel/Privacy-Preserving-ML/, this repository offers transparency and reproducibility, enabling fellow researchers and practitioners to validate the results, explore the methodologies employed, and build upon the findings to advance the field of privacy-preserving machine learning.

A.2 Real-World Examples of PPML

Project Title	*Objective*	*Approach*	*Parameters*	*Privacy Budget*
Apple QuickType Suggestions [1]	Learn previously-unknown words typed by sufficiently many users	ϵ-DP	$\epsilon = 8$	$\epsilon = 16$ collected per day
Apple Emoji Suggestions [1]	Calculate which emojis are most popular among users	ϵ-DP	$\epsilon = 4$	$\epsilon = 4$ collected per day
Facebook Full URLs Dataset [49]	Provide data on user interactions with web pages shared on Facebook	(ϵ, δ)-DP	$\epsilon = 0.45$, $\delta = 10^{-5}$	Not Specified
Facebook Movement Range Maps [33]	Quantity the changes in mobility of Facebook users during the COVID-19 pandemic	ϵ-DP	$\epsilon = 1$	$\epsilon = 2$ over the two statistics collected
Google Gboard Next Word Prediction [47]	Predict the next word to be typed by a user	(ϵ, δ)-DP	$\epsilon = 6.92$, $\delta = 10^{-5}$	Not Specified
Google Urban Mobility Data [13]	Provide data on whether users traveled from one location to another during a given time period	(ϵ, δ)-DP	$\epsilon = 0.66$, $\delta = 2.1 \cdot 10^{-29}$	Not Specified
LinkedIn Audience Engagement API [55]	Allow marketers to get information about LinkedIn users engaging with their content	(ϵ, δ)-DP	$\epsilon = 0.15$, $\delta = 10^{-10}$	$\epsilon = 34.9$, $\delta = 7 \cdot 10^{-9}$ per month
Microsoft Global Victim Perpetrator Synthetic Dataset [6]	Provide data about victims and perpetrators of trafficking	(ϵ, δ)-DP	$\epsilon = 12$, $\delta = 5.8 \cdot 10^{-6}$	Not Specified
US Census Post-Secondary Employment Outcomes [27]	Provide data about the earnings and employment of college graduates	ϵ-DP	$\epsilon = 1.5$	$\epsilon = 3$ over the two statistics collected

Table 8. Real-World Applications of Differential Privacy [19]

A.3 Experiment Results

A.3.1 Differential Privacy - Regression (Student Performance Dataset).

A.3.2 Differential Privacy - Classification (Census Income Dataset).

A.3.3 Differential Privacy - Regression (Household Power).

A.3.4 Federated Learning + Differential Privacy Time Efficiency (Household Power).

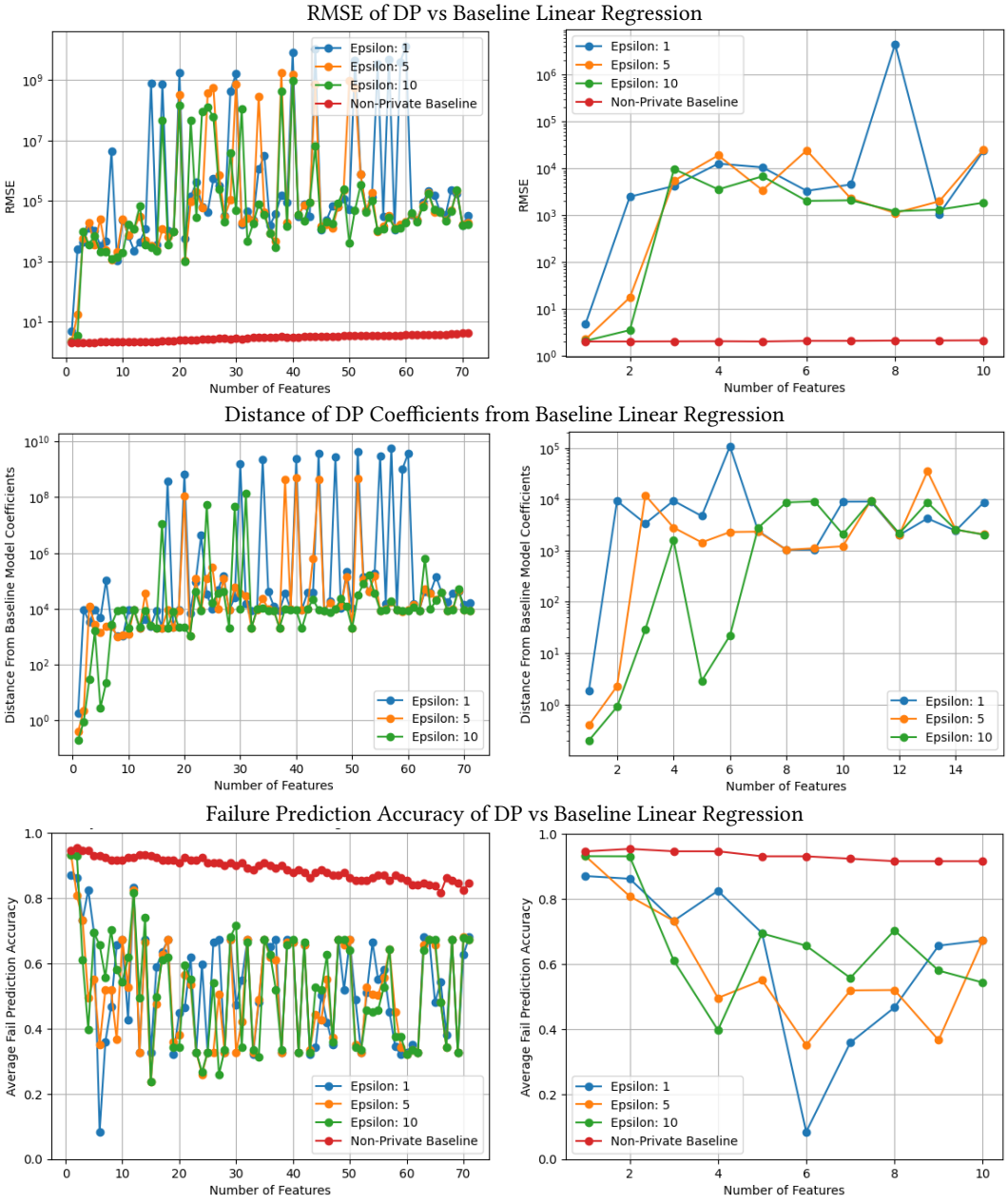


Table 9. DP vs Baseline Linear Regression Model for Different Feature Set Sizes (Student Performance Dataset)

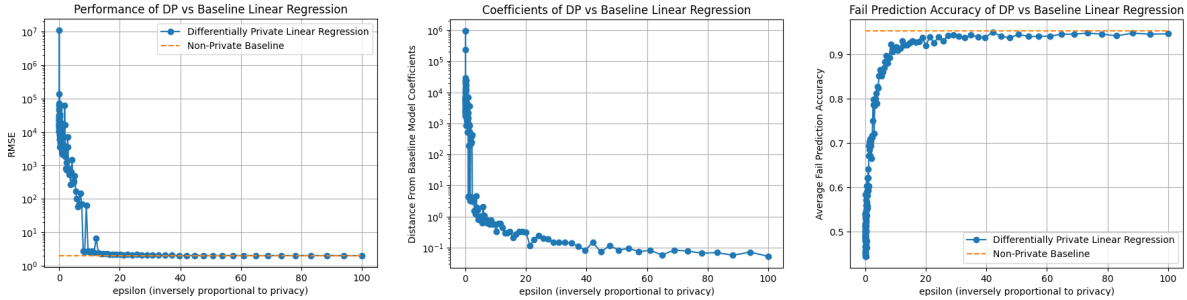


Table 10. DP vs Baseline Linear Regression Model using Top 2 Features (Student Performance Dataset)

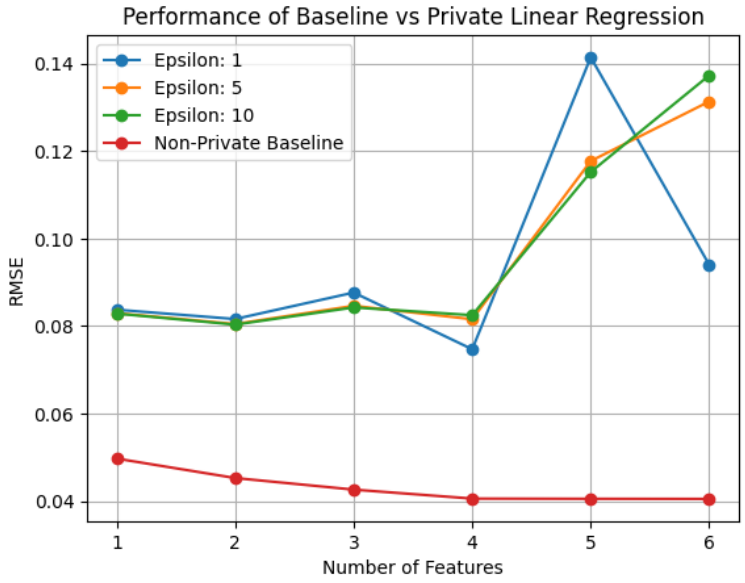
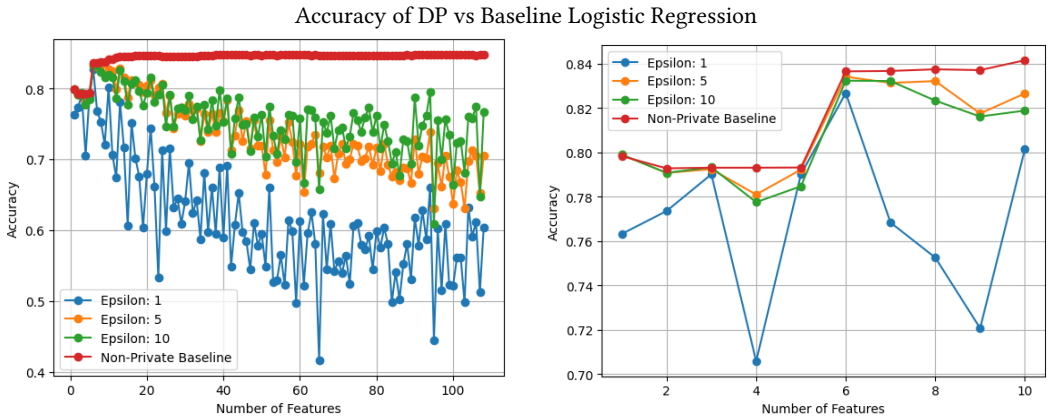
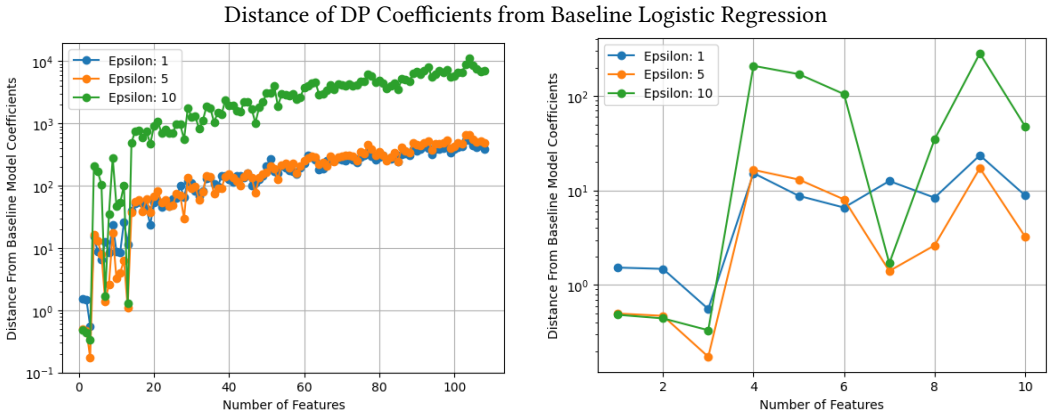


Fig. 9. Time Efficiency of FL-DP vs Baseline Model (Household Power Dataset)





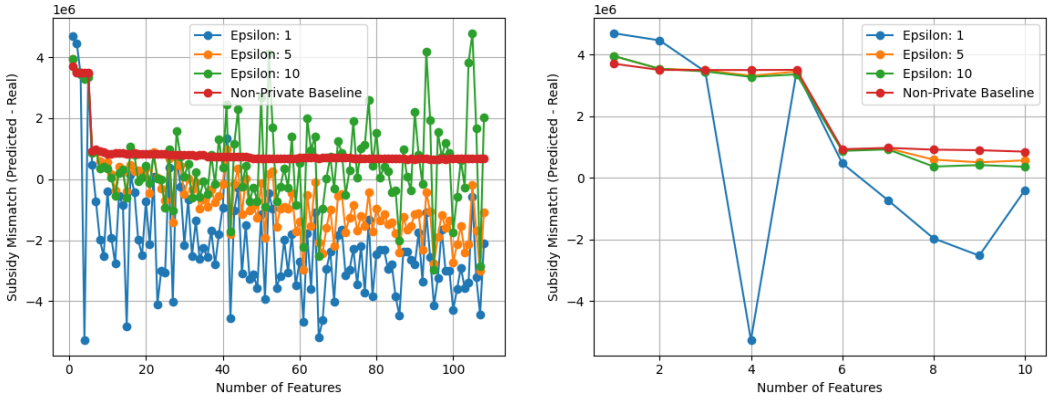


Table 11. DP vs Baseline Logistic Regression Model for Different Feature Set Sizes (Census Income Dataset)

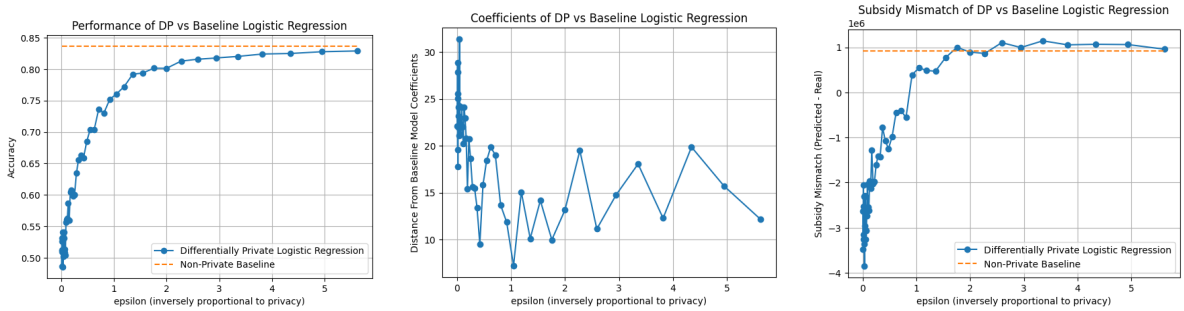


Table 12. DP vs Baseline Logistic Regression Model using Top 6 Features (Census Income Dataset)

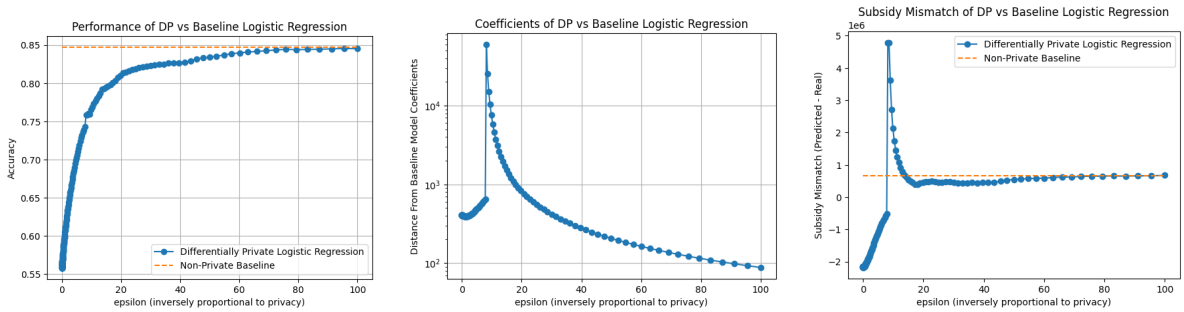


Table 13. DP vs Baseline Logistic Regression Model using All 108 Features (Census Income Dataset)

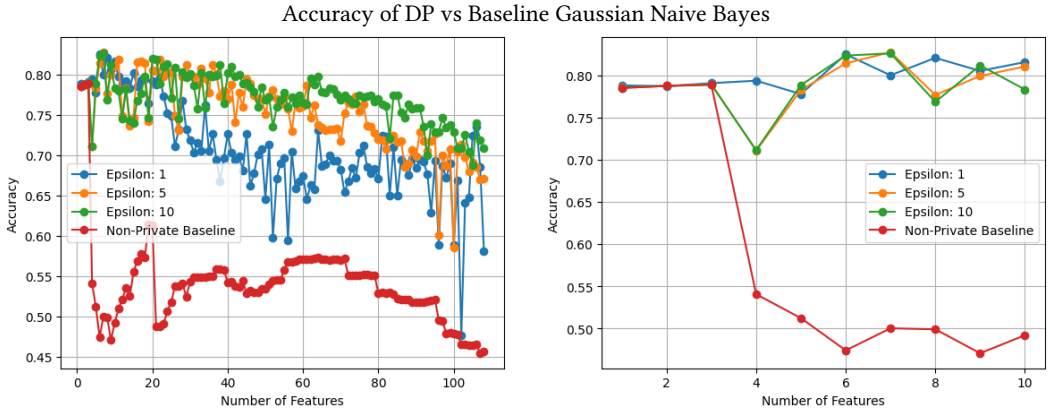


Table 14. DP vs Baseline Gaussian Naive Bayes Model for Different Feature Set Sizes (Census Income Dataset)

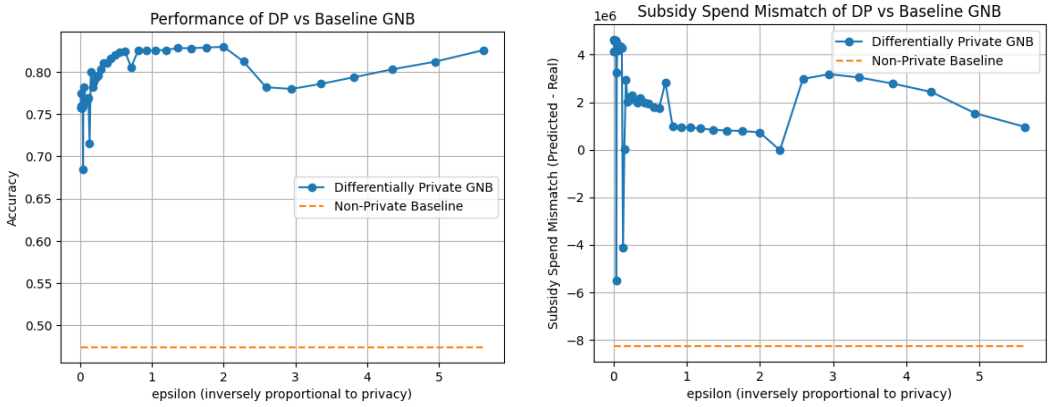


Table 15. DP vs Baseline Gaussian Naive Bayes Model using Top 6 Features (Census Income Dataset)

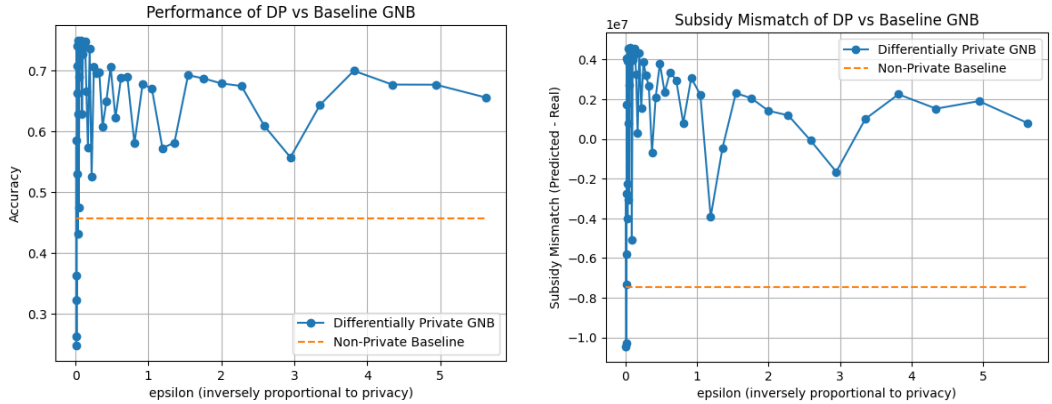
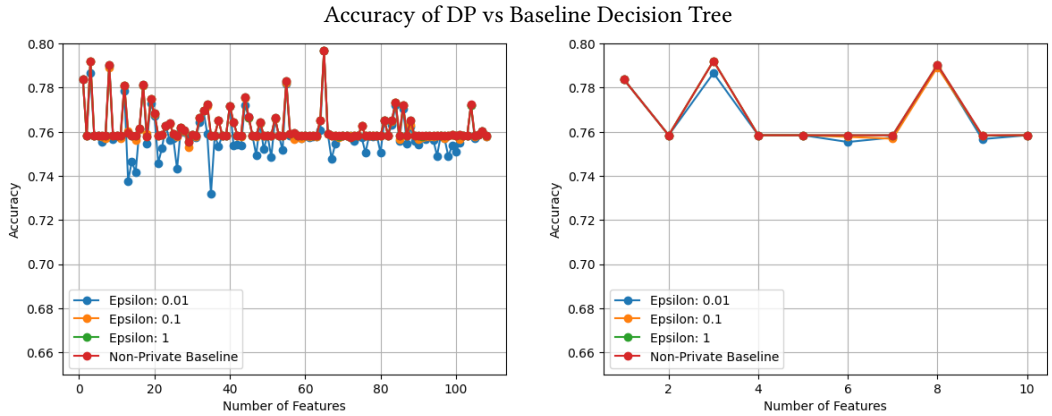


Table 16. DP vs Baseline Gaussian Naive Bayes Model using All 108 Features (Census Income Dataset)



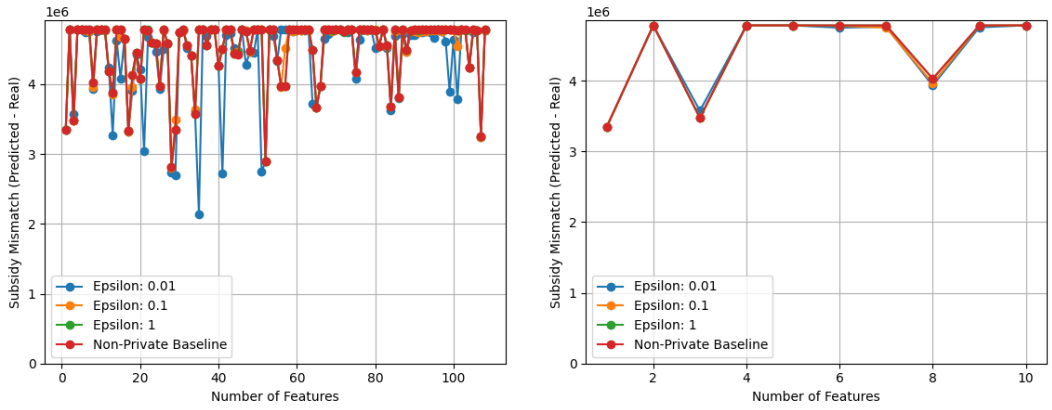


Table 17. DP vs Baseline Decision Tree Model for Different Feature Set Sizes (Census Income Dataset)

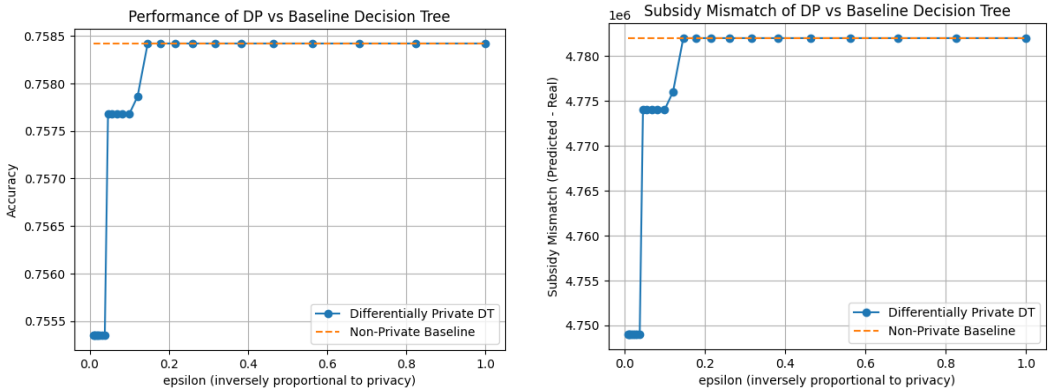


Table 18. DP vs Baseline Decision Tree Model using Top 6 Features (Census Income Dataset)

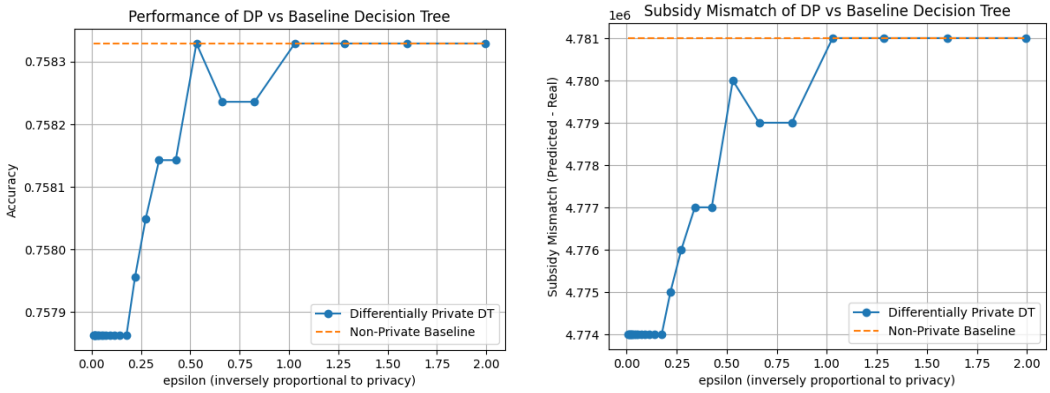


Table 19. DP vs Baseline Decision Tree Model using All 108 Features (Census Income Dataset)

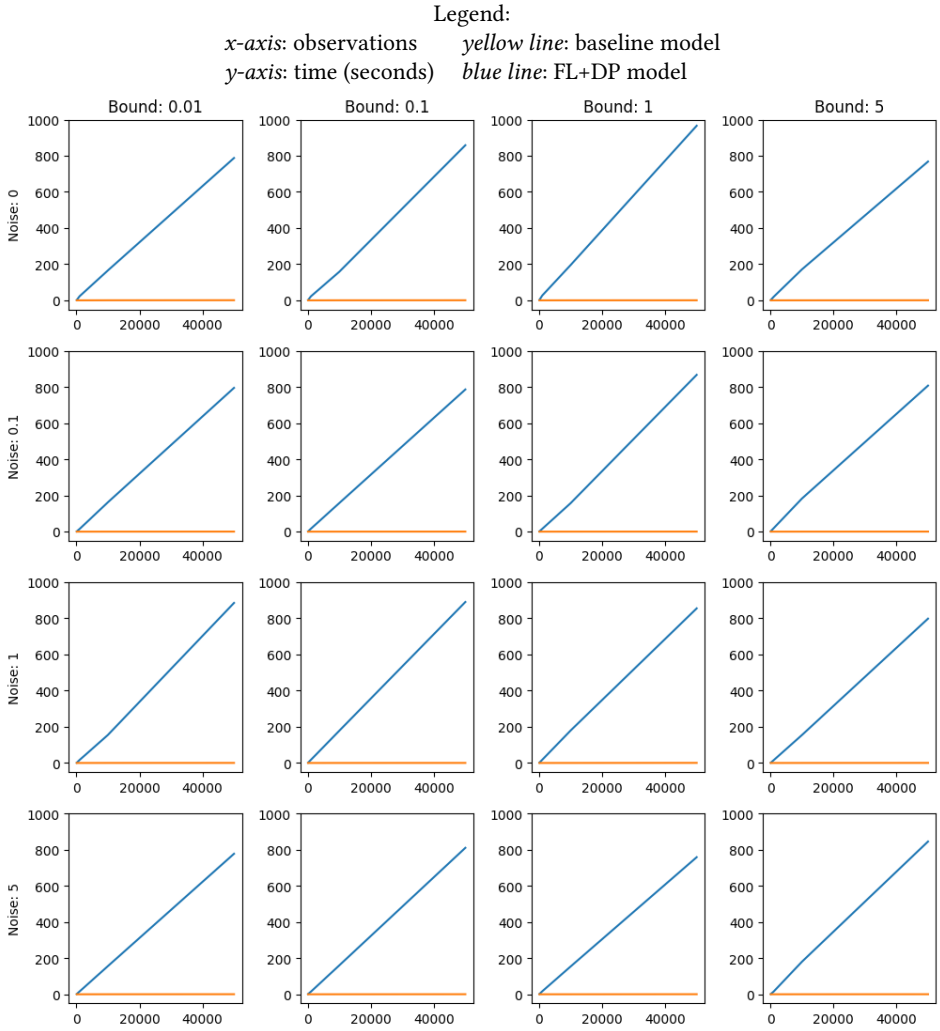


Fig. 10. Time Efficiency of FL-DP vs Baseline Model (Household Power Dataset)