Open Peer Review on Qeios

Secure and Private Machine Learning: A Survey of Techniques and Applications

Raja Abou[1]

1 Isik University

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

Machine Learning (ML) privacy violations can lead to discrimination and identity theft, among other serious repercussions. As more sensitive data has been utilized to train models in recent years, the necessity for privacy-preserving methods in ML has grown in significance. The state-of-the-art methods for Privacy-Preserving Machine Learning (PPML), such as safe multi-party computation, homomorphic encryption, and differential privacy, are thoroughly reviewed in this survey study. We also assess PPML's drawbacks, including scalability, computing efficiency, and the trade-off between privacy and utility. Finally, we identify the open problems and future directions of research in PPML, including emerging trends, challenges, and opportunities. This survey paper is intended to serve as a valuable resource for researchers and practitioners interested in the area of PPML.

Mhd Raja Abou Harb

Computer Engineering, Faculty of Engineering and Natural Sciences, Işık University, Istanbul, Türkiye (email address: 21comp9001@isik.edu.tr)

Keywords: Privacy-preserving machine learning, Homomorphic encryption, Machine learning, Privacy violation, Differential privacy, Safe multi-party computation.

1. Introduction

In today's life, computing power interacts with almost all of our life parts. This high dependence on computers can lead to serious privacy violations. One of the major trends in computer usage is ML. Based on (Burkov, 2019), ML is "a field of study that focuses on the design and development of algorithms that can learn from and make predictions on data, without being explicitly programmed. In other words, it involves the development of computational models that can automatically improve their performance on a given task, based on the feedback they receive from the data.". It is being used in solving a lot of life problems in many fields such as healthcare, finance, and many more. This technique uses datasets that contain features to build the needed knowledge to have correct answers. There is a claim that states the more data records are used in the training phase the higher accuracy that can be get in the testing phase (Hey, Tansley, & Tolle, 2009). From that point of view, scientists started to request data related to their research fields to apply them to ML models. If we take medical care as an example and based on Health Insurance Portability and Accountability Act (HIPAA) regulations, this information shall safeguard the privacy and security of individuals' health information by establishing national standards for the use and disclosure of Protected Health Information (PHI) by covered entities and business associates. When these data are being shared with scientists for their experiments, a lot of countermeasures are being taken into consideration, such as hiding Personal Identity Information (PII), however, the main question is: are these countermeasures enough for protecting PII and PHI data?

Several researchers thought about that question; (Kim, Kim, & Kim, 2017) found that there is a possibility of de-anonymizing medical data in South Korea that include Resident Registration Numbers (RRNs), which serve as distinctive identifiers for every Korean resident, examined in this research. The authors show that RRNs can be re-identified with a high degree of accuracy using information that is readily accessible to the public. Moreover, (Sweeney, 2002) claimed that three bits of information—the person's date of birth, gender, and Zone Improvement Plan (ZIP) code—could be used to uniquely identify 87% of the United States (US) population. Using data anonymization techniques for protecting PII is not enough. And from that point of view, the PPML domain had been raised.

PPML is the technique of developing ML algorithms and models that safeguard sensitive data privacy during the inference and training phases.

In traditional ML, models are frequently trained using data that has been gathered, consolidated, and used for a variety of applications. When sensitive data, like financial or personal health records, is used, this method, however, may give rise to privacy concerns. By creating methods that allow for the deployment and training of ML models without jeopardizing the privacy of the underlying data, PPML tackles this problem.

PPML approaches can be divided roughly into two groups:

1. Differential privacy. This method conceals users' identities and stops the leakage of private information by introducing random noise to the data.
2. Secure Multiparty Computation (SMC). Multiple parties can cooperatively train an ML model without disclosing any of their private data. The data is exchanged and encrypted in a way that enables each party to do calculations on the data without having direct access to it.

In this work, a novel overview of PPML is going to be discussed. This review can help the readers in finding the research gaps and suggest future work related to this field.

The organization of this paper will be according to the following: background terms that can help in understanding the topic that is going to be discussed in Section 2. Section 3 will discuss the used techniques for PPML. Section 4 will discuss the newest papers related to this topic. Section 5 the challenges of this topic, and finally section 6 is the conclusion.

2. Background

In this section foundational concepts that are needed are going to be mentioned. The terms in this section can be useful in understanding the concepts of the research topic that we are studying.

2.1. ML models

ML is considered part of the Artificial Intelligence (AI) domain. It tries to build a way to help the machine to make decisions in a lot of problems. ML algorithms can be classified into 2 major domains: Supervised and unsupervised ML algorithms. The main difference between the two categories is having labels from the experts of the domain that is studied. If the algorithm needs in the learning process a dataset with labels, the algorithm is considered supervised ML. In general, ML algorithms can detect the patterns that are available in the features of the used datasets, and based on that they can make predictions of new cases.

2.2. Encryption in PPML

By facilitating the safe exchange, storage, and processing of private data, cryptography is essential for PPML. To encrypt the data and carry out computations on it without disclosing its contents to unauthorized parties, cryptographic techniques such as homomorphic encryption, safe multiparty computing, and secret sharing are used. Homomorphic encryption allows computations to be performed on encrypted data without first decrypting it, ensuring that sensitive information is not revealed during the computation. Secure multiparty computation allows multiple parties to jointly compute a function on their data without revealing their data to each other. Secret sharing splits the data into several pieces that are distributed among different parties in such a way that only authorized parties can access and reconstruct the original data. These cryptographic techniques enable the development of PPML algorithms that can protect the privacy of sensitive data while still enabling the extraction of useful insights and knowledge. Figure 1 shows an overview of homomorphic encryption in the ML process (Olzak, 2022).

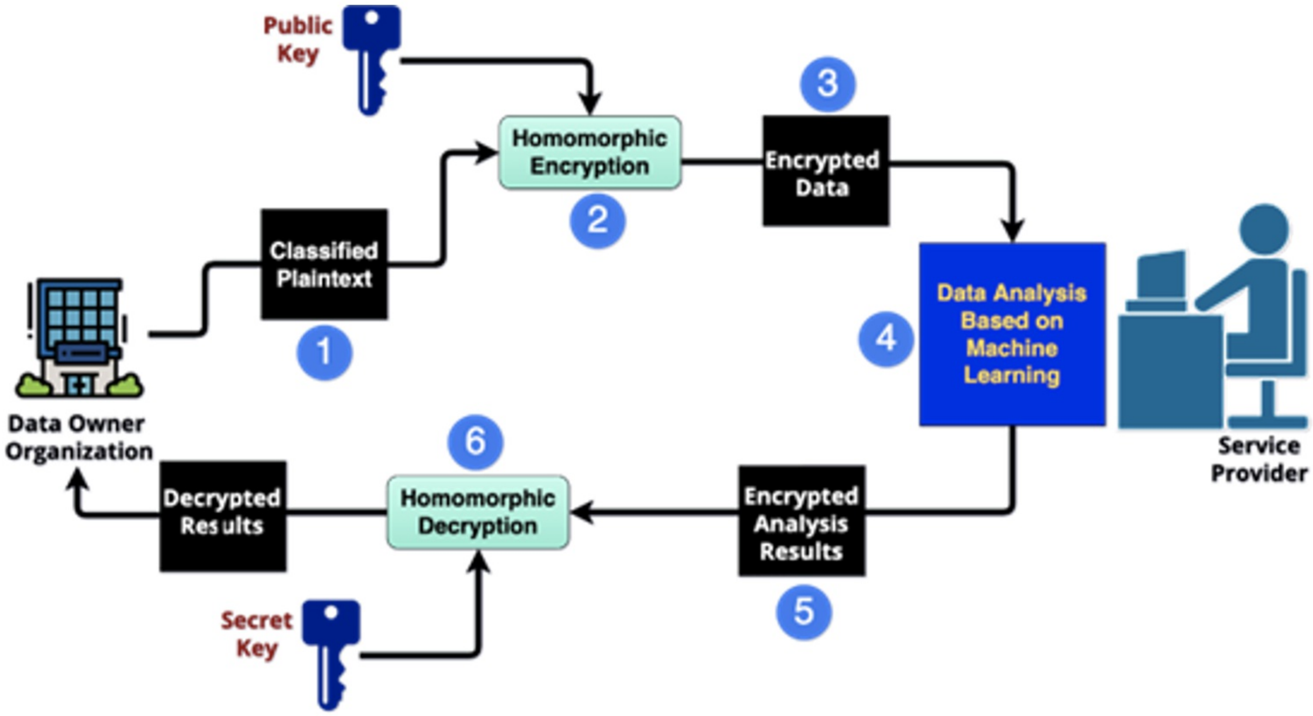


Figure 1. Homomorphic Encryption In ML Models.

2.3. Deferential Privacy

Deferential privacy offers a precise specification of privacy guarantees in the context of data analysis. It makes sure that no private information about any dataset participant is revealed in the analysis's results.

The concepts of "privacy loss" and "privacy budget" can be used to explain the idea of differentiated privacy. When a single person's data is added to or deleted from the dataset, the output of the mechanism does not dramatically alter, the mechanism is said to be differentially private.

Mathematically, a randomized algorithm or mechanism is said to satisfy ε-differential privacy if, for any pair of datasets D and D' that differ by the inclusion or exclusion of a single individual's data, and for any set of possible outputs S, the following inequality holds:

$$\Pr\left[M(D) \in S\right] \leq \exp(\varepsilon) \Pr\left[M(D') \in S\right]$$

In this equation, $\Pr\left[M(D) \in S\right]$denotes the probability that the mechanism M applied to dataset D outputs a result in the set S. The ε parameter controls the amount of privacy protection provided, where a smaller ε value implies stronger privacy guarantees.

To illustrate this with an example, consider a scenario where a company wants to release aggregate statistics about the average income of its employees while preserving the privacy of individuals. Applying differential privacy, the company can add random noise to the true average income in a controlled manner.

Suppose the true average income of employees is 50,000 TL. Without differential privacy, an attacker might be able to determine the salary of an individual employee by comparing the released average income with external information. However, by applying differential privacy, the company adds random noise to the average income. For example, the reported average income could be 50,200 TL or 49,800 TL with equal probability. This additional randomness ensures that an attacker cannot reliably infer the exact income of any individual employee based on the released statistics.

By carefully controlling the amount of noise added (which is determined by the ε parameter), differential privacy allows for a balance between preserving individual privacy and providing accurate aggregate information. A smaller ε value provides stronger privacy guarantees but may introduce more noise, potentially reducing the accuracy of the released data, while a larger ε value may provide less privacy but yield more accurate results. For more information about deferential privacy, please refer to (Dwork, 2006)

3. Privacy-preserving for ML techniques

Given that the models and algorithms used in ML frequently depend on sizable volumes of sensitive data, privacy is a crucial challenge in PPML. Creating ML models that can learn from sensitive data without revealing any sensitive information about the people who provided the data is the aim of PPML. This is especially crucial when dealing with confidential information that should be kept secret, such as financial or personal health data. Several privacy protection strategies are frequently employed in ML to ensure privacy. Data anonymization, encryption, and access restriction are some of these mechanisms. Before using data for ML, personally identifying information is subtracted from it in a process known as data anonymization. Removing names, addresses, and other data that could be used to identify specific people is one way to do this. Data is encoded during encryption so that only those with the proper permissions can access it. To restrict who can access the data and how access control measures are utilized. These controls are essential for ensuring that private information is protected and sensitive data is kept private throughout the ML process.

In addition to these tools, PPML requires careful thought over the data processing and sharing protocols applied during model inference and training. The way these protocols are designed is essential for preventing the leakage or improper use of sensitive data. To prevent the re-identification of people, differential privacy might be employed, for instance, to amplify the data. Multiple parties can collaboratively compute a function on their unique data using secure multiparty computing without disclosing their personal information to one another. Data is divided into many pieces and transferred among several parties using secret sharing, making it possible for only authorized parties to access and reassemble the original data. These cryptographic methods make it possible to create ML algorithms that can safeguard the privacy of sensitive data while yet allowing for the extraction of insightful information. To preserve confidence in the algorithms and to make sure that people are not at risk of having their private information compromised, privacy protection in ML is essential.

4. Related works in PPML

In this section, the efforts in the academic work in PPML are going to be discussed. We chose the important research papers in this domain to study how they worked on solving the research problem.

4.1. (Al-Rubaie, 2019)

This article can be considered a good resource to start on the PPML topic. The authors provided an overview of the threats and challenges to privacy in ML, as well as the solutions and techniques for preserving privacy during the training and inference phases of ML models.

Firstly, they discussed the importance of applying extra countermeasures to increase privacy in ML models. They showed several scenarios of possible unauthorized access or misuse of shared data for ML models. Based on their classification three actors can involve in the ML process: data owners, processors, and the people who are going to receive the results. The highest security can be achieved when these actors are related to the same party or person, however, not all cases can be like that. They mentioned five levels of threats to privacy leaks in ML models. These threats are cleartext of private data, reconstruction attacks, model inversion attacks, membership inference attacks, and Re-Identification. In the cleartext of private data, the data is stored in storage in plaintext when it is transferred to the computation part. This may affect privacy especially when the data storage is compromised or internal threats by disloyal employees. Experts suggested sending features that are needed instead of sending the whole row of data, however, this solution leads to the second type of threat which is reconstruction attacks. Having the knowledge or metadata of the extracted features can help in reconstructing the row data which can lead to privacy violations. The attacker in a model reversion assault can produce feature vectors that mimic those used to build an ML model by using the output from that ML model. These attacks make use of the confidence data which is sent back as a response. Knowing if one of the member's or participant's data had been used in the training phase can lead us to the fourth type of thread which is membership inference attacks. One of the solutions that were proposed for preserving the privacy of participated individuals in building datasets was to remove any feature or column that can be related to PII, however, based on the fifth threat, which is Re-Identification, attackers can reach these omitted data by combining the collected dataset with other datasets.

The study then discusses several methods and solutions for protecting privacy in ML, including federated learning, homomorphic encryption, secure multi-party computation, and differential privacy. Each of these methods is thoroughly detailed, along with the underlying ideas, benefits, and potential applications.

The problems and trade-offs of PPML are also covered in the study, along with the ethical and legal issues that come up when managing sensitive data. These issues include the impact on model accuracy, computation time, and communication overhead.

The paper offers a thorough and understandable summary of the state-of-the-art in PPML and emphasizes the need for additional research and development in this field to satisfy the changing risks and needs for privacy in ML.

4.2. (Liu, Guo, Lam, & Zhao, 2022)

The authors in this article proposed a scalable privacy-preserving scheme, which is tolerant to drop-out any participant at any time. they used in their proposal Homomorphic encryption Pseudorandom Generator (HPRG) and Shamir secret sharing scheme. The proposed solution can reduce the communication overhead by avoiding the need to construct seeds used for generating masks from the client side. There is also no need to send additional Shamir shares to servers in case of dropped clients similar to the SecAgg-based schemes that they used in their comparison. They claim that their proposal is stronger than other solutions for dropout resilience.

In the security analysis of the solution, four theories and their proof were mentioned. Two of these theories talked about the protection level of the system from semi-honest parties. In their definition, semi-honest participants “will not deviate from the protocol but try to infer the honest parties’ information.”. the first theorem concerns the semi-honest participants excluding the server side. For all U, t, k with $/C/<t, x_U, U_1, U_2, U_3$and C where $C \subseteq U$ and $U_3 \subseteq U_2 \subseteq U_1$ there exists a probabilistic polynomial-time (PPT) simulator SIM such that

$$SIM_C^{U,t,k}\left(x_U, U_1, U_2, U_3\right) \equiv REAL_C^{U,t,k}\left(x_U, U_1, U_2, U_3\right)$$

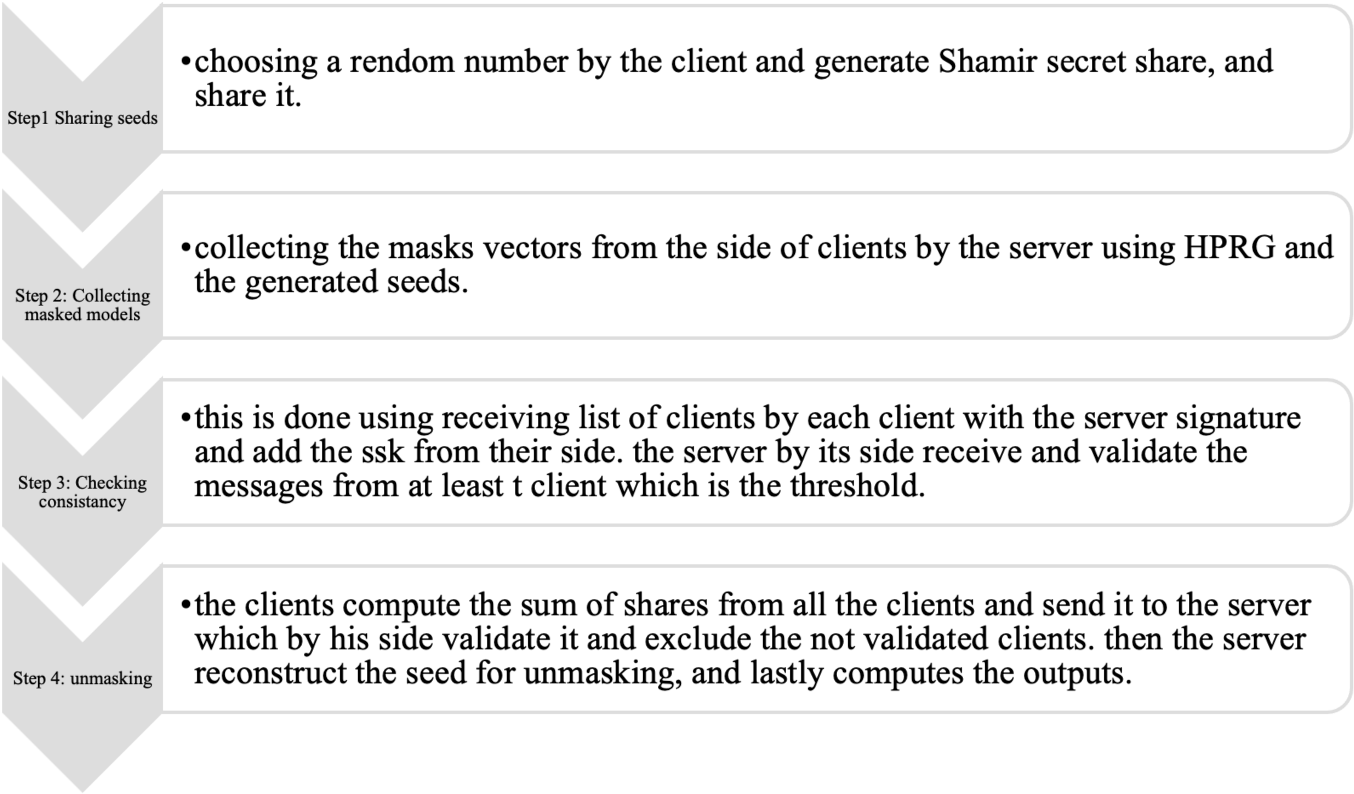
The second theory is concerned with the security in case of semi-honest participants including the server. For allU, t, k with $/C\backslash\{S\}/<t, x_U, U_1, U_2, U_3$and C where $C \subseteq U \cup S$ and $U_3 \subseteq U_2 \subseteq U_1 \subseteq U$ there exists a probabilistic polynomial-time (PPT) simulator SIM such that

$$SIM_C^{U,t,k}\left(x_U, U_1, U_2, U_3\right) \equiv REAL_C^{U,t,k}\left(x_U, U_1, U_2, U_3\right)$$

The other two theories talked about security against active malicious clients. To provide the needed security, a new HPRG-based Secure aggregation protocol was proposed. The third theory concerns malicious clients and honest servers. For all U, t, k with $/C/<t, x_{U\backslash C}$, and C with the algorithm M_C , where $C \subseteq U$, the proposed protocol is a secure protocol for computing $F_{HSecAgg}$, and the last theory concerns with malicious clients include the server.

To understand the theories, we need to know the denotations of the terms inside it$U$ indicates a set of clients with their locally trained models x_U . *S indicates the server side of the model.*The equality in both theories indicates the identity in distributions for both sides.

As an overview of the proposed protocol, the private inputs can be summarized by the locally trained model or gradient for each client, which can be presented as vectors, secrets for initiating secret keys, and signature secret keys. From the server side, secret keys for authenticating secret channels and for signatures are stored. On the other hand, the public inputs for the protocol are the number of clients, the threshold of dropped clients, and the public keys for signature processes. The aggregation results of locally trained models can be considered the output of the protocol. Figure 2 shows the protocol’s steps and a brief description of each step.



Step1 Sharing seeds

- choosing a rendom number by the client and generate Shamir secret share, and share it.

Step 2: Collecting masked models

- collecting the masks vectors from the side of clients by the server using HPRG and the generated seeds.

Step 3: Checking consistancy

- this is done using receiving list of clients by each client with the server signature and add the ssk from their side. the server by its side receive and validate the messages from at least t client which is the threshold.

Step 4: unmasking

- the clients compute the sum of shares from all the clients and send it to the server which by his side validate it and exclude the not validated clients. then the server reconstruct the seed for unmasking, and lastly computes the outputs.

Figure 2. HPRG-based Secure Aggregation Protocol Steps.

Several tests proceeded in the article which proved the assumptions that they mentioned.

4.3. (Gupta & Singh, 2022)

In this article, the authors worked on proposing a cloud-based PPML model using differential privacy and ML model. The proposed model specifies communication protocol between untrusted parties. The conducted experiments included Naïve Bayes (NB) classifier and it was applied to several datasets., and they reached 95% of accuracy which was higher than the studied literature. Figure 3 shows the proposed model which is called the Differential Approach for data and classification service-based Privacy-preserving Machine Learning Model (DA-PMLM).

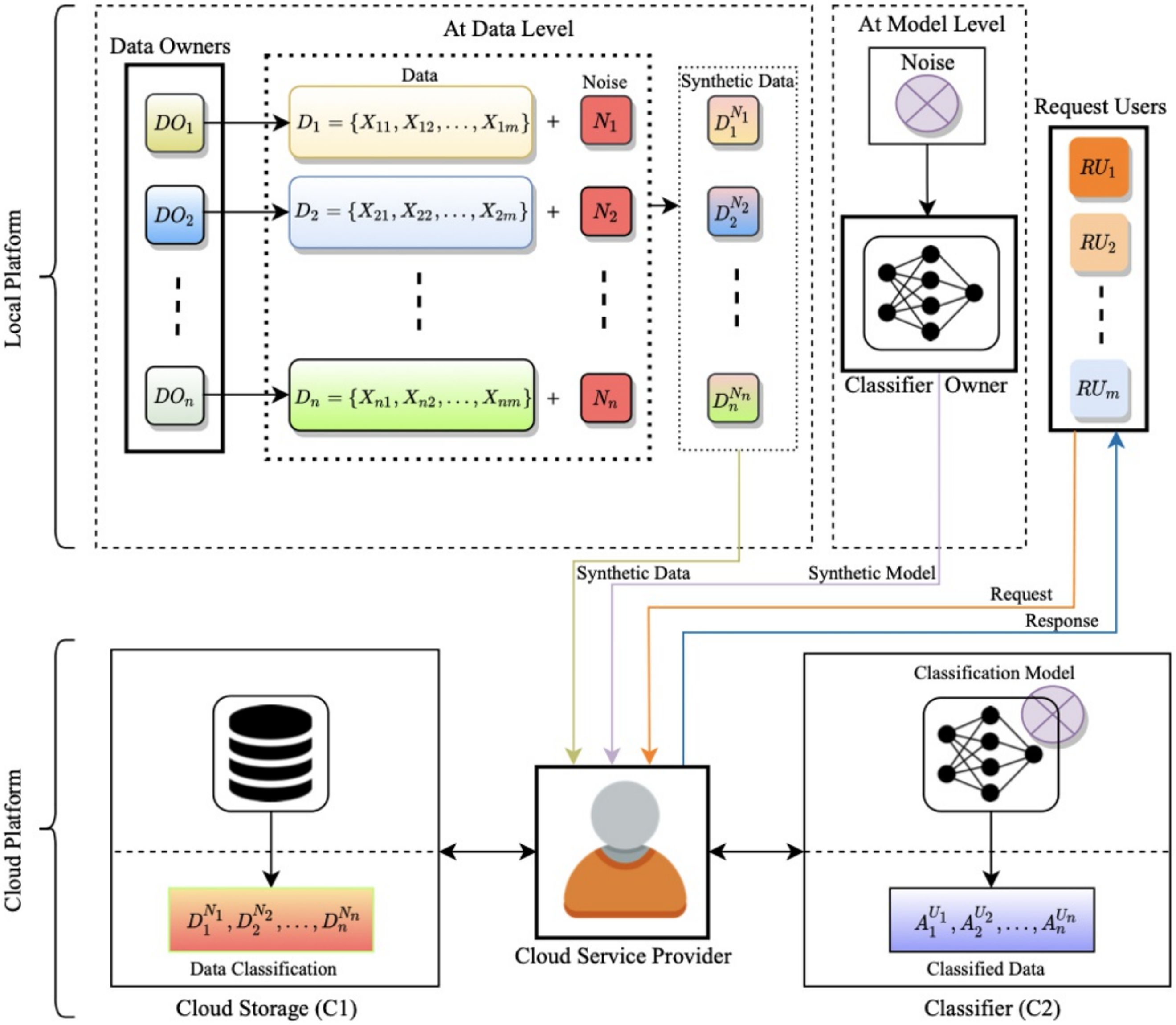


Figure 3. Proposed Architecture.

From Figure 3, four main actors can be found in the architecture; are Data Owners (DO_{id}) which can be called cloud client, who is the creator of the data and willing to share it on the cloud, Classifier Owner (CO) which is the party which concerns with providing classification services to the Cloud Service Provider (CSP) which by its side provides services to DO_{id}, and finally the Request Users (RU_{id}) who are the entities who request owner's data from CSP. DOs inject noise into their private data before sharing it using differential privacy.

They proofed the theory which states "In the proposed model, the privacy-preserving mechanism of the classification model satisfies the parallel composition of ε-differential privacy"

Figure 4 shows the classification and data flow of the proposed model.

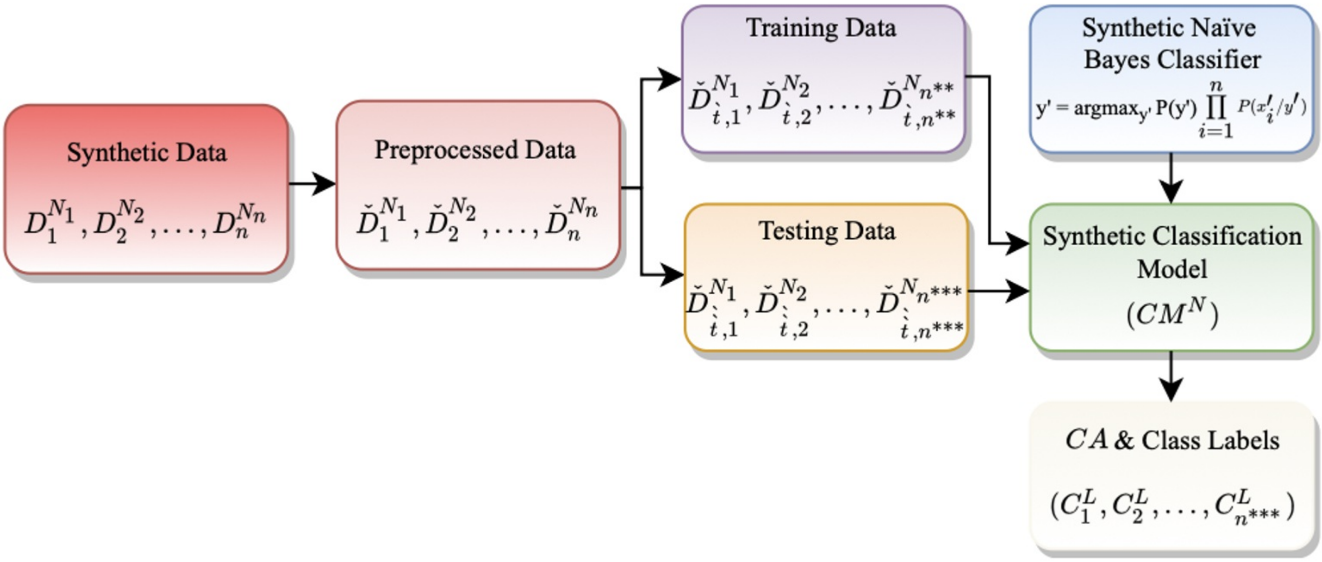

Synthetic Data
$D_1^{N_1}, D_2^{N_2}, \ldots, D_n^{N_n}$
Preprocessed Data
$\check{D}_1^{N_1}, \check{D}_2^{N_2}, \ldots, \check{D}_n^{N_n}$
Training Data
$\check{D}_{t,1}^{N_1}, \check{D}_{t,2}^{N_2}, \ldots, \check{D}_{t,n^{**}}^{N_{n^{**}}}$
Testing Data
$\check{D}_{t,1}^{N_1}, \check{D}_{t,2}^{N_2}, \ldots, \check{D}_{t,n^{***}}^{N_{n^{***}}}$
Synthetic Naïve Bayes Classifier
$y' = \text{argmax}_{y'} P(y') \prod_{i=1}^{n} P(x_i'/y')$
Synthetic Classification Model
(CM^N)
CA & Class Labels
$(C_1^L, C_2^L, \ldots, C_{n^{***}}^L)$

Figure 4. The Classification and Data flaw proposed Solution.

After receiving the data with noise D_1^{N1}, D_2^{N2}, …D_n^{Nn} from the corresponding Dos, CSP preprocesses them by applying normalization functions using Z-score normalization equation. The resulting values can be denoted using the symbol $\check{D}_i^{Ni}$. Similar to any ML process the given dataset is divided into training and testing parts. After that, the training and testing of the Classification Model (CM), which is in the experiment NB classifier proceeded to see the evaluation measurements.

4.4. (Lee, et al., 2022)

The authors of this article claim that the current PPML solutions are limited to non-standard ML models. These solutions are not proven good results in practical datasets. Moreover, the used activation functions are simple from the arithmetic point of view and replaced non-arithmetic activation functions. They also avoid using bootstrapping. A large number of layers is not possible in the current solutions too, their proposal includes implementation of the standard ResNet-20 model with the RNS-CKKS Fully Homomorphic Encryption (FHE) with bootstrapping. To verify their model, they applied it to a standardized dataset which is CIFAR-10. The reached results were highly nearby the usage of the original ResNet-20 model with no encryption, and the reached accuracy was about 92.43%.

5. Challenges in PPML

PPML faces several challenges that hinder its widespread adoption in practical applications. One of the primary challenges is the trade-off between privacy and utility. PPML techniques such as differential privacy and secure multi-party computation often introduce noise or communication overhead, which can significantly impact the accuracy and efficiency of ML models. Another challenge is the lack of standardization and interoperability among different PPML techniques, making it difficult to compare and combine results from different approaches. Additionally, PPML requires specialized expertise and resources, such as knowledge of cryptography and secure computation, which may not be readily available to all organizations. Finally, PPML raises legal and ethical issues, such as regulatory compliance and transparency, that must be carefully addressed to ensure the responsible and trustworthy use of sensitive data.

6. Conclusion and future work

The privacy of sensitive data must be preserved, and this field of study on PPML is one that is actively and quickly evolving. The significance of privacy protection systems cannot be stressed as ML becomes more widely used in industries like healthcare, banking, and social media. To ensure that people are not at risk of having their sensitive information revealed, privacy-preserving algorithms and protocols that can enable ML while respecting privacy are essential. Progress in this area of research, which encompasses several branches of computer science including information theory, ML, and cryptography, depends on multidisciplinary cooperation. PPML is a challenging and complex problem, but one that is necessary to ensure that the benefits of ML can be realized without sacrificing the privacy of individuals. We have outlined some of the primary difficulties and upcoming research topics in this survey study, as well as given an overview of the main methods and techniques utilized in PPML. In order to overcome the significant issues in this field, we hope that this survey study will be a valuable resource for researchers and practitioners who are interested in PPML.

For future work, more studies about PPML on real data shall be done and comparing the performance with classical ML models.

Bibliography

- Sweeney, L. (2002). k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 557-570.
- Burkov, A. (2019). *The Hundred-Page Machine Learning Book.*
- Dwork, C. B. (2006). Part II, chapter Differential Privacy. *CALP 2006.* Springer.
- Al-Rubaie, M. &. (2019). Privacy-preserving machine learning: Threats and solutions. *IEEE Security & Privacy*, 49-58.
- Liu, Z., Guo, J., Lam, Y., & Zhao, J. (2022). *Efficient dropout-resilient aggregation for privacy-preserving machine learning.* IEEE Transactions on Information Forensics and Security.
- Lee, J. K., Lee, E., Lee, J., Yoo, D., Kim, Y., & No, J. (2022). Privacy-preserving machine learning with fully homomorphic encryption for deep neural network. *IEEE Access*, 30039-30054.
- Gupta, R., & Singh, A. K. (2022). A differential approach for data and classification service-based privacy-preserving machine learning model in cloud environment. *New Generation Computing.*
- Olzak, T. (2022, July 6). *Homomorphic Encryption: How It Changes the Way We Protect Data* Retrieved from Spiceworks: https://www.spiceworks.com/it-security/data-security/articles/how-homomorphic-encryption-protects-data/
- Hey, T., Tansley, S., & Tolle, K. (2009). *The Data Deluge: An eScience Perspective.* Taylor & Francis Group.
- Kim, S., Kim, J., & Kim, J. (2017). De-anonymizing South Korean Resident Registration Numbers Shared in

Prescription Medical Records. *Healthcare Informatics Research*, 138-144.