PRIVATE-AI: A hybrid approach to privacy preserving AI

Saif Kamalsha
Computer Science & Engineering
PES University
Bangalore,India
saif.kamalsha@gmail.com

Siri S
Computer Science & Engineering
PES University
Bangalore,India
sirisrinivasag@gmail.com

Siddanth Krishna
Computer Science & Engineering
PES University
Bangalore,India
sid.ullal@gmail.com

Sai Amruth
Computer Science & Engineering
PES University
Bangalore,India
saiamruth3@gmail.com

Shruti Jadon
Computer Science & Engineering
PES University
Bangalore,India
shrutijadon@pes.edu

***Abstract* — Machine learning algorithms based on Deep Neural Networks (NN) have achieved remarkable results and are being extensively used in different domains. A crucial enabler for Artificial Intelligence is the quantum of data, and these models are as good as the quality of data that is used to train them. However, gathering data and using them to prognosticate behaviors presents great challenges to the privacy of individuals and organizations, such as data breaches, privacy loss, and the corresponding financial and reputational damages. Privacy-preserving machine learning (PPML) aims at bridging the gap between preserving privacy and reaping the benefits of ML. It is a key enabler for privatizing collected data and complying with data protection regulations.**

Therefore, the goal of our project is to implement a hybrid approach to Privacy Preserving AI by using these three Privacy Preserving techniques namely: Federated Learning, Differential Privacy, and Homomorphic Encryption to achieve maximum user data privacy while not affecting the overall accuracy of the predictions and computations made by the AI model.

Keywords—federated learning , differential privacy, secure aggregation using homomorphic encryption

I. Introduction

Applications for machine learning and artificial intelligence have expanded rapidly in the past ten years. In fact, AI is the new electricity that is revolutionizing virtually every business. Data and computing power are the engines propelling this degree of growth. For the foreseeable future, both the amount of data produced and processing capacity will continue to increase dramatically. We can create better AI models if we have access to more data because the results of AI models are only as good as the data. The need for additional data is evident, but we also require data that is sufficiently diverse and updated often.

Large databases owned by organisations include our personal information. They utilise this information to develop forecasts specifically for us. For instance, Netflix uses our personal information to make personalised movie suggestions, Amazon uses it to make product recommendations, and Facebook uses it to make advertising recommendations. And let's face it, we are all dependent on these tailored experiences. Let's consider the following dilemma: Should we compromise our privacy for personalization?

Data scientists typically gather and assemble data in one place and use it to train ML models. However, because so much of the data in the world is intended to remain private, researchers and technologists have been working to create solutions that are not reliant on a single data source. This concept serves as the cornerstone of privacy-preserving machine learning systems, also referred to as federated learning (FL).

The machine learning model is locally trained at the source in a federated environment, which could be a data source or an edge device with private user data. The central model is then updated using the locally trained model, which is then delivered back to the central site. Therefore, ML models that allow the data to remain in its original place while certain ML model information is transferred between locations can be built using federated learning. Information that has been exchanged does not intentionally divulge sensitive or private information.

Differential privacy is a mathematical strategy that prevents anyone from learning information about the individuals in a dataset by introducing a predetermined amount of randomness to the dataset. Controlled randomness has been added. As a result, the final dataset still contains enough reliable data to produce overall insights while protecting the privacy of individual individuals. Due to this, differential privacy allows businesses to tailor the privacy level and exposes attackers to incomplete data.

Homomorphic encryption is a type of encryption that has the ability to do additional evaluations on encrypted data without having access to the secret key. Such a calculation yields an encrypted result. One might think of homomorphic encryption as a development of public-key cryptography. Homomorphic encryption consists of a number of encryption methods that allow for various classes of computations over encrypted data. The computations are represented by either Boolean or arithmetic

circuits. Some popular types of encryption include leveled fully homomorphic, totally homomorphic, substantially homomorphic, and slightly homomorphic encryption.

In order to achieve maximum user data privacy without compromising the overall accuracy of the predictions and computations made by the AI model, the project aims to implement a hybrid approach to privacy-preserving AI using these three privacy-preserving techniques: federated learning, differential privacy, and homomorphic encryption.

The rest of this dissertation is divided into the following sections:

II. RELATED WORK

McMahan *et al.* [13] illustrate the application of FL; however, Malekzadeh et al. [12], Truex et al. [1], and Wei et al. [3] improved it by including DP in the accumulation process.

The first to propose a hybrid solution to this problem was Truex et al. [1]. For regional training and secure federated aggregation carried out at a central aggregator, they employ a number of clients. To update the central model, weights sent from the client to the centre are homomorphically encrypted, averaged, and then decrypted using the majority of the individual clients' keys. Similar to how Malekzadeh et al. [12] design this hybrid approach, they consider the trade-off between employing momentum during training and several Local Stochastic Gradient Descent (SGD) stages in [1].

Some hybrid methods that combine FL and HE to securely train models are those by Stripelis et al. [7] and Sav et al. [8]. However, in [8], the entire model is encrypted, and all training is performed using the model's encrypted parameters, which is computationally expensive. To get around this, [7] uses unencrypted weights for local training and only encrypts the weights when they need to be transferred to a centralised controller for aggregation. Even yet, since each client has access to the aggregated un-noisy (without DP) model weights in plaintext, it does not shield against model inversion attacks in the case of client compromise.

In the inference phase, Dowlin et al. [9], Chou et al. [10], and Disabato et al. [11] have offered effective methods for using HE, including the replacement of operations that HE does not support with polynomial approximations [9]. All of the known methods either safeguard an individual's privacy while building a model [1], [1], [3], [8], or while using the model to make predictions [9]–(11), with just a few methods describing privacy protection in both Scenarios simultaneously.

III. PROPOSED APPROACH

We identify 3 key issues that exist in the realm of privacy, namely, (1) User Data Privacy: centralised data centers pose a threat to clients who send across their private data. (2) Training Data Privacy: malicious actors can reverse engineer vulnerable user data. (3) Parameter privacy: the central model has access to key weight updates.

In order to tackle the aforementioned issues, we list three separate Privacy Preserving techniques

A. Federated Learning

Federated learning is a method of training machine learning models on decentralized data that is distributed across multiple devices or locations, without the need to centralize the data. This allows organizations to train machine learning models on data that is distributed across a network of devices, such as smartphones or internet of things (IoT) devices, without the need to collect and centralize the data on a single server.

Federated learning has several benefits compared to traditional machine learning approaches that require centralized data. For example, it allows organizations to train machine learning models on a much larger and more diverse dataset than would be possible with a centralized approach, and it allows organizations to preserve the privacy of individual data points by training the model on the device without revealing the data to a centralized server. Federated learning typically involves the following steps:

- Identify the devices that will participate in the federated learning process.
- Select the data to use for training on each device.
- Train a local model on each device using the selected data.
- Aggregate the local models to create a global model.
- Use the global model to make predictions on new data.

Federated learning can be applied to a wide range of machine learning tasks, including image recognition, natural language processing, and recommendation systems. It has been used in many different applications, such as improving the performance of mobile keyboard applications, optimizing energy usage in smart buildings, and enhancing the accuracy of medical image analysis.

B. Differential Privacy

Differential privacy is a technique for protecting the privacy of individuals in a dataset by adding carefully designed noise to the data before it is used for training or analysis. This allows organizations to perform machine learning or other data-intensive tasks on sensitive data, while ensuring that the privacy of individual data points is preserved.

Differential privacy adds noise to the data in a way that satisfies a mathematical definition of privacy, known as the "privacy budget". This definition specifies the maximum amount of information that can be leaked about an individual data point, without violating the privacy of that data point. The privacy budget is determined by the level of noise added to the data, and it can be adjusted to trade off between privacy protection and accuracy.

Differential privacy has been used in many different applications, such as training machine learning models on medical data, analyzing sensitive survey data, and improving the security of search engines. It has been shown to be effective at protecting the privacy of individual data points while still allowing organizations to gain insights from the data.

We intend to use Differential Privacy in our approach to guarantee that the training data is safe from being exposed/reverse engineered.

C. Homomorphic Encryption

Homomorphic encryption is a type of encryption that allows mathematical operations to be performed on ciphertext, that is

encrypted data, in a way that is consistent with the operations performed on the plaintext, or unencrypted data. This means that it is possible to perform operations on encrypted data, without the need to decrypt the data first.

Homomorphic encryption has several advantages over traditional encryption methods. For example, it allows organizations to perform operations on sensitive data without revealing the data to unauthorized parties, and it allows organizations to share encrypted data with other parties without the need for them to have the decryption key. This makes it useful for applications where data privacy and security are important, such as in the healthcare and finance industries.

Homomorphic encryption is a relatively new and active area of research in the field of cryptography, and there are many different approaches and techniques that have been developed for implementing homomorphic encryption. Some of the most common types of homomorphic encryption include partially homomorphic encryption, fully homomorphic encryption, and somewhat homomorphic encryption. Each of these approaches has its own strengths and limitations, and they are used in different applications depending on the specific requirements of the system.

We intend to use HE in our combined approach as it protects the trained model from being learned by any external parties including the server with minimal to no accuracy loss since it does not add any noise.

D. Training

i. Preprocessing Data

The MNIST dataset is a widely-used benchmark dataset for image classification tasks in machine learning. The images in this dataset are handwritten digits and come in a standard format of 28x28 pixels. However, before training a machine learning model on this dataset, it is often necessary to preprocess the data to improve the model's accuracy. One common preprocessing step is to apply transforms to normalize the MNIST data. This involves applying mathematical operations to scale and shift the pixel values of the images so that they have a mean of zero and a standard deviation of one. Normalizing the data in this way can help to improve the model's performance by ensuring that the data has a consistent range of values and reducing the impact of outliers. Additionally, by normalizing the data, the model can learn the important features of the data more efficiently. Some commonly used transforms for normalizing MNIST data include mean normalization, standard deviation normalization, and scaling. Overall, applying transforms to normalize the MNIST dataset is an important preprocessing step that can significantly improve the performance of machine learning models on this dataset.

ii. Creation of Virtual Workers and Use of DP

Federated learning is a privacy-preserving technique that allows multiple parties to collaboratively train a machine learning model on their decentralized data sources without having to share their raw data. To facilitate this collaborative training process, virtual workers can be created. These virtual workers are software agents that mimic the behavior of real workers, allowing for distributed training while maintaining data privacy. Each virtual worker is assigned a subset of the data and performs computations locally on that data. The results are then aggregated to create a global model, which is then shared back to each virtual worker. By using virtual workers in federated learning, the privacy of the data sources is protected, while also enabling the creation of a more robust and accurate model. Virtual workers can also be used to control the level of participation in the federated learning process, allowing for dynamic allocation of computational resources to ensure efficient and effective training. Differential Privacy is to be added to these inputs by these virtual workers to the local ML models to ensure better privacy levels in the model. Overall, the use of virtual workers is an important aspect of federated learning, enabling secure and collaborative training without compromising the privacy of the data sources.

iii. Federated Averaging with Encrypted Parameters

To train models on virtual workers without moving any gradients to the central model until the gradients have been collated, a technique called "Federated Averaging" can be used. In this method, each virtual worker computes a gradient update based on its local dataset, and these updates are sent to the central server. In this approach however, the parameters that are expected to be aggregated are encrypted via Homomorphic Encryption (HE) to protect the parameters exiting the local models to produce more security for the workers' data. However, the server does not update the global model with these gradients immediately. Instead, the gradients are collected from all the virtual workers and averaged to create a global gradient update. This averaged update is then applied to the global model, and the process is repeated for the next round of training. This technique ensures that the gradients remain local and are never moved to the central server until they have been aggregated with other gradients, reducing the risk of exposing sensitive data. Additionally, by averaging the gradients, the final model is more robust and generalizable, as it is trained on a larger and more diverse dataset.

E. Inference

Before training a machine learning model on datasets such as MNIST, it is often necessary to preprocess the data to improve the model's accuracy. One common preprocessing step is to apply transforms to normalize the data, such as mean normalization and scaling. Normalizing the data can help to improve the model's performance by ensuring that the data has a consistent range of values and reducing the impact of outliers.

To achieve the privacy proposed by this model, virtual workers are to be created. To train models on these virtual workers, federated averaging can be used, where gradients are aggregated before being sent to the central server, ensuring that gradients remain local until they have been collated. Once received, the server aggregates the updates and computes a new set of model parameters before sending them back to each remote model to update their local models. Overall, by using privacy-preserving techniques in federated learning and applying appropriate preprocessing steps, machine learning models can be trained efficiently and accurately on decentralized datasets while preserving the privacy of the individual data sources.

IV. IMPLEMENTATION

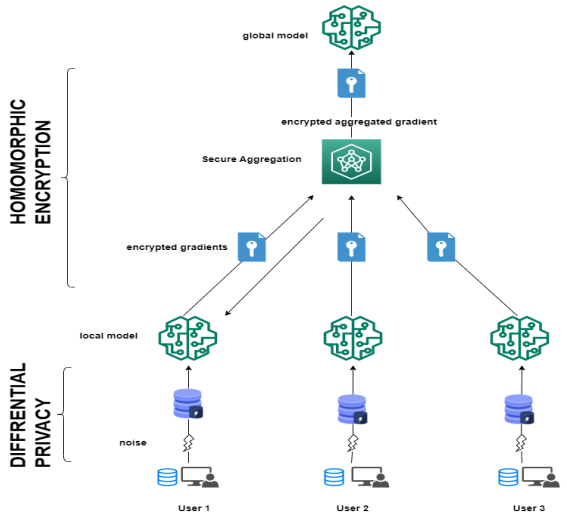
The above mentioned components of privacy preservation is incorporated in the said hybrid approach :

- Firstly, in Federated Learning (FL), the model itself is sent across to the end users who perform computation locally.

Then the updated weights are sent back to the global server which aggregates it.

- Secondly, in Differential Privacy (DP), noise is added to each client. This is done to ensure that no reconstruction or re-identification attacks can be performed by malicious entities.

- Lastly, in Homomorphic Encryption (HE), both encryption and decryption are done locally at the client using public and private HE keys. This increases privacy and accuracy.



A. Convolution Neural Networks

From the Bottom-Up: First, the users parameters are taken and the model adds noise to it (differential privacy), then our model encrypts and securely sends across these parameters to the aggregator (homomorphic encryption) this is combined with a Federated Learning Architecture as user data never leaves local nodes.

V. EXPERIMENTAL RESULTS

A. Centralised Machine Learning

It was determined how well vanilla machine learning performed before and after the addition of differential privacy and homomorphic encryption in phases.

The findings displayed in the table below show that vanilla ML has outstanding accuracy. Accuracy decreases when Differential Privacy is included to the aforementioned ML model. The privacy metric, however, indicates an improvement in security. 0 (theoretically) represents the maximum level of privacy and infinite the lowest level of privacy in the differential privacy scale, which ranges from 0 to infinity.

TABLE I: Accuracy of various techniques

Technique	Accuracy	Inference
ML only	0.971	Accuracy is brilliant
ML + Differential Privacy	0.92 (epsilon=0.563)	Accuracy reduces slightly, however privacy increases
ML + Homomorphic Encryption	0.95	Accuracy is good, however takes time to predict

B. Comparison Metrics of Centralised and Decentralised Learning

The difference in accuracy between a vanilla machine learning model (centralised learning) and the presence or absence of federated learning (decentralised approach) was noted. Both IID (Independent and identically distributed) data and Non-IID data were used to generate the metrics. Figure 1 demonstrates how Federated Learning has improved accuracy for both IID and Non-IID data.

The conclusion drawn from the Table 2 is further explained in the table below. The inclusion of federated learning has clearly increased accuracy. Even more private than standard centralised learning, federated learning offers.

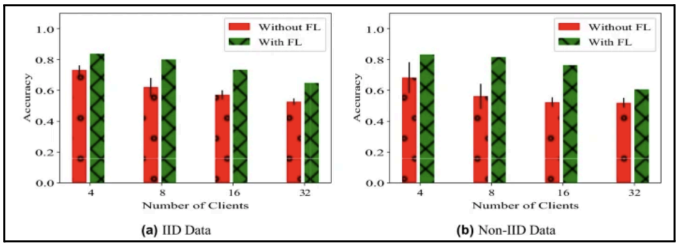


Fig 1: Comparison of vanilla ML with FL based on accuracy against an increase in clients

Data distribution	Number of clients *n*	Accuracy		
		Without FL	With FL	Centralized
IID	4	0.731 ± 0.03	0.824 ± 0.02	0.848 ± 0.02
	8	0.620 ± 0.06	0.780 ± 0.05	
	16	0.570 ± 0.03	0.726 ± 0.06	
	32	0.527 ± 0.02	0.641 ± 0.09	
Non IID	4	0.682 ± 0.10	0.824 ± 0.01	0.848 ± 0.02
	8	0.561 ± 0.08	0.823 ± 0.05	
	16	0.524 ± 0.03	0.750 ± 0.06	
	32	0.520 ± 0.03	0.550 ± 0.20	

Table 3: Comparison of accuracy with and without Federated Learning on IID and Non-IID data

C. Comparison Metrics of Federated learning with Differential Privacy

TECHNIQUE	ACCURACY	INFERENCE
FL only	0.924	Privacy improves significantly.
FL + DP (Noise multiplier = 4, Grad Clipping = 1)	0.815 and ε = 2.90	The accuracy decreases with the intro of DP but privacy increases.

Table 4: Comparison of accuracy with and without Differential Privacy on decentralised learning

Testing was done to compare the accuracy and privacy of plain federated learning against adding differential privacy. Due to the inclusion of noise to the data, the adoption of Differential Privacy reduces accuracy while significantly increasing privacy. Reverse engineering attempts to recover the original data are prevented by the insertion of noise. Also, because client data is never transmitted to the server in federated learning, privacy is considerably increased.

5.4 Our Approach - PRIVATE AI

This is the intended strategy when the three approaches covered in this paper—federated learning, differential privacy, and homomorphic encryption—are combined.

The training data is kept local to their respective trusted owners to guarantee privacy, and the model parameters are differentially private to avoid indirect inference. Only the client themself is able to access the client data in plaintext for inference. Without the proper keys, which are exclusively available at the client, the encrypted data cannot be decrypted or inferred even if the Global Server or other parties have access to it. The client will never be aware of the classification model that was employed because the model stays at the global server. The results clearly indicate a significant increase in the privacy metric of the environment when PRiVATE AI (see results in Table 5), our suggested approach to counteract the threat of data leakage in AI systems is implemented as opposed to any other feasible combination of Privacy Preserving techniques with a slight loss in model accuracy.

TECHNIQUE	ACCURACY (PRIVACY LOSS)	INFERENCE
FL + Secured Aggregation (HE)	0.937	Accuracy does not change drastically but time taken to train increases significantly

Table 5: Comparison of accuracy with and without Differential Privacy on Federated learning + Secured Aggregation (HE)

VI. CONCLUSION

There are several benefits to using privacy preserving AI systems. For example, they allow organizations to use sensitive data for AI applications without violating the privacy of individuals, and they can improve the security and trustworthiness of AI systems by reducing the risk of data breaches or unauthorized access to sensitive data. Additionally, privacy preserving AI systems can enable organizations to train and operate machine learning models on decentralized data, which can improve the performance and scalability of the AI system.

According to the results obtained in the implementation of the desired hybrid approach, it can be inferred that this is a private and secure approach that can be implemented on machine learning models.

Privacy has significantly increased in comparison to a vanilla federated learning approach or a combination of federated learning with either differential privacy or homomorphic encryption. Accuracy has not increased much in comparison to a centralised machine learning technique.

The setup used for training the model is based on a real world scenario where client data is dispersed among several users in an untrustworthy environment. Therefore, the users do not share data amongst themselves to train the ML model. This problem is solved by Differential Privacy and Federated Learning. Federated learning helps keep the data localized to a user, and Differential Privacy prevents reverse engineering attacks that are performed by third parties. Homomorphic Encryption comes into play to protect the privacy of clients from attack vectors and the central server itself during transit.

This work will increase the confidence of users to use their data in the field of artificial intelligence without the worry of security. Even though computational complexity is a drawback due to HE, given better memory and compute resources, this can be used on a larger scale across various users.

In summary, privacy preserving AI systems are better than traditional AI systems because they provide the benefits of AI technology without compromising the privacy of individuals.

VII. FUTURE WORK

By using advanced techniques to protect sensitive data, these systems can improve the security, trustworthiness, and performance of AI systems, while still providing valuable insights and predictions from the data.

This novel approach is an open area of research and can be extended in multiple directions in the future in terms of implementing it on models other than CNN, providing encryption at every level and trying to use secure multiparty computation with public key cryptography similar to [7], [8] approaches to enhance protection of data in transit over the network. The other area that can be looked into is the accuracy of the approach.

All these improvements can be used to guarantee better privacy of individuals even though they might increase the computational cost of individuals and negatively affect accuracy.

VIII. ACKNOWLEDGEMENTS

The authors would like to thank the Computer Science and Engineering Department of PES University for providing the opportunity and resources to complete this work.

IX. REFERENCES

[1] Stacey Truex, Thomas Steinke “A Hybrid Approach to Privacy-Preserving Federated Learning” Georgia Institute of Technology, August 2019. (references)

[2] Georgios A. Kaissis, Marcus R. Makowski “Secure, privacy-preserving and federated machine learning in medical imaging” Nature Machine Intelligence, August 2019. (references)

[3] Kang Wei, Jun Li, Ming Ding, Chuan Ma “Federated Learning with Differential Privacy: Algorithms and Performance Analysis” ARXIV, Nov 2019

[4] W. Kim and J. Seok, "Privacy-preserving collaborative machine learning in biomedical applications," 2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC).

[5] Takabi, Hassan and Ehsan Hesamifard. “Preserving Multi-party Machine Learning with Homomorphic Encryption.” (2016).

[6] H. Zhang, J. Bosch and H. H. Olsson, "Federated Learning Systems: Architecture Alternatives," 2020 27th Asia-Pacific Software Engineering Conference (APSEC), 2020

[7] Stripelis, D., Saleem, H., Ghai, T., Dhinagar, N., Gupta, U., Anasta-siou, C., Ver Steeg, G., Ravi, S., Thompson, P. & Ambite, J. Secure Neuroimaging Analysis using Federated Learning with Homomorphic Encryption. (2021,8)

[8] Sav, S., Pyrgelis, A., Troncoso-Pastoriza, J., Froelicher, D., Bossuat, J., Sousa, J. & Hubaux, J. POSEIDON: Privacy-Preserving Federated Neural Network Learning. (2021)

[9] Dowlin, N., Gilad-Bachrach, R., Laine, K., Lauter, K., Naehrig, M. & Wernsing, J. CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy. Proceedings Of The 33rd International Conference On International Conference On Machine Learning - Volume 48. pp. 201-210 (2016)

[10] Hesamifard, E., Takabi, H. & Ghasemi, M. CryptoDL: Deep Neural Networks over Encrypted Data. (2017)

[11] Disabato, S., Falcetta, A., Mongelluzzo, A. & Roveri, M. A PrivacyPreserving Distributed Architecture for Deep-Learning-as-a-Service. 2020 International Joint Conference On Neural Networks (IJCNN). pp. 1-8 (2020)

[12] Malekzadeh, M., Hasircioglu, B., Mital, N., Katarya, K., Ozfatura, M. & Gündüz, D. Dopamine: Differentially Private Federated Learning on Medical Data. (2021)

[13] McMahan, H., Moore, E., Ramage, D., Hampson, S. & Arcas, B. Communication-Efficient Learning of Deep Networks from Decentralized Data. (2017)