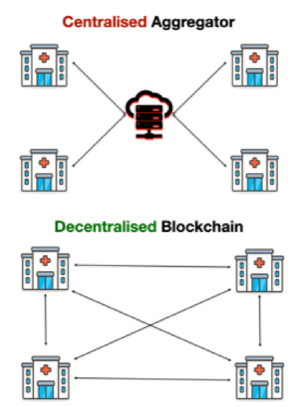
BEAS: Blockchain Enabled Asynchronous & Secure Federated Machine Learning

Harpreet Virk   Arup Mondal   Debayan Gupta

Ashoka University

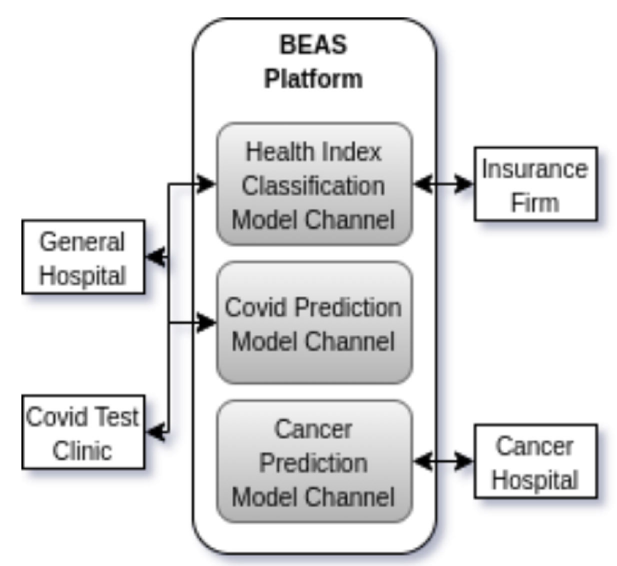
Motivation

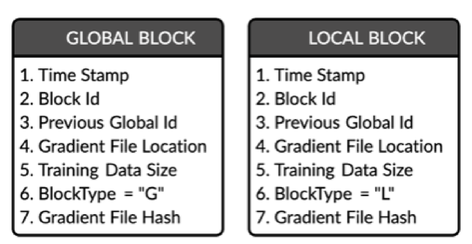
Federated Learning (FL) assumes trust in the centralized aggregator which stores and aggregates model updates. These shared gradients are susceptible to various inference attacks that can leak sensitive information. They are also vulnerable to adversarial poisoning attacks.



BEAS Framework

Beas aims to achieve secure and efficient N-party ML while ensuring strict privacy guarantees using **Gradient Pruning** based differential privacy, and resiliency from poisoning attacks using **FoolsGold** and **Multi-KRUM**. With approximately 92.72% accuracy on MNIST, Beas achieves training accuracy comparable with both – centralized and non-privacy preserving decentralized approaches.





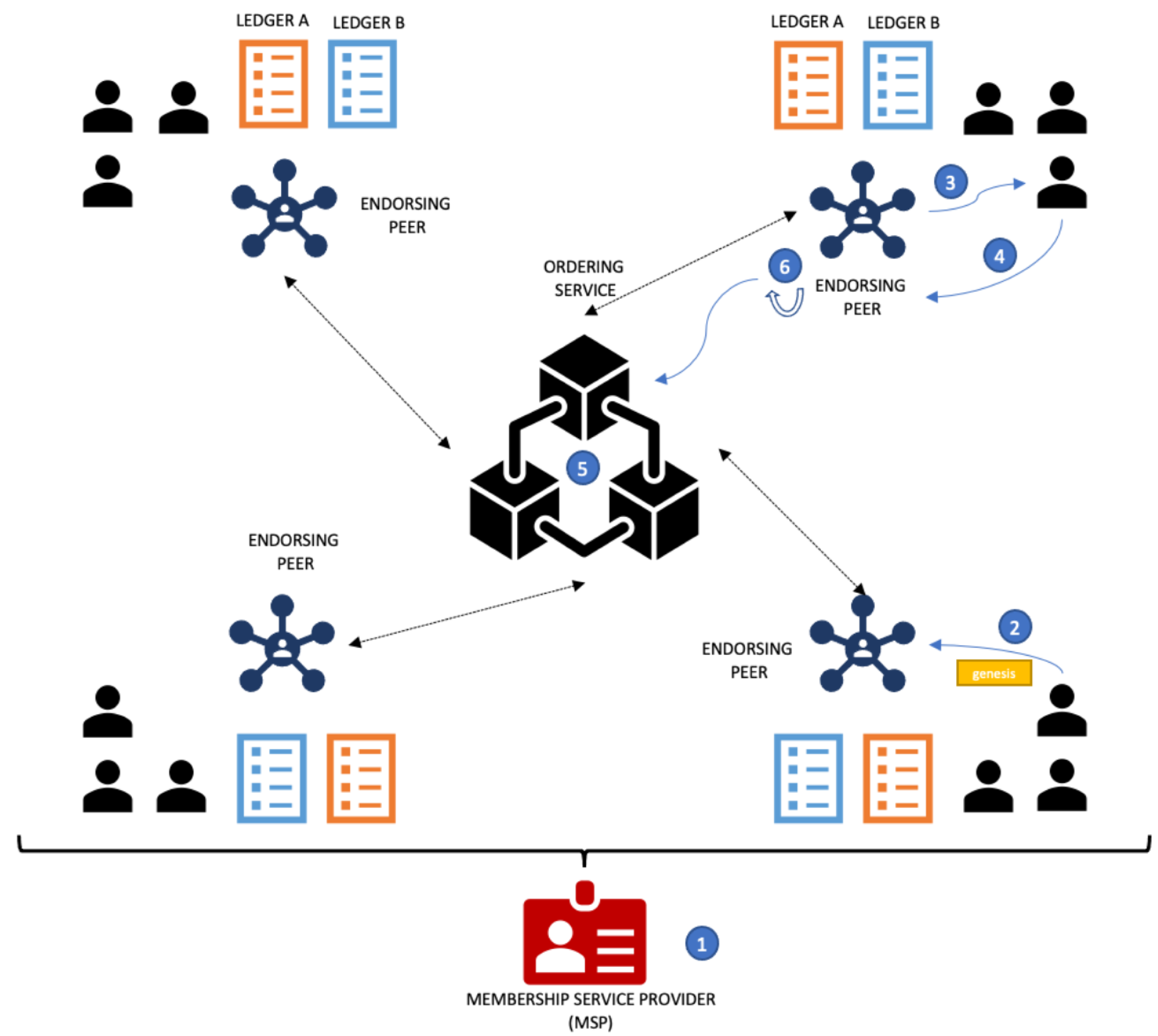
Dataset	Centralized Accuracy (%)	$N = 20$ Accuracy (%)	$N = 20$ Avg. Execution Time (s) Local Training	$N = 20$ Avg. Execution Time (s) Overall	$N = 50$ Accuracy (%)	$N = 50$ Avg. Execution Time (s) Local Training	$N = 50$ Avg. Execution Time (s) Overall
MNIST	95.53	92.74	1.89	524	90.11	1.89	726
Malaria	98.89	96.16	2.18	967	92.81	2.18	1276
CIFAR-10	72.81	61.03	38	21608	63.76	38	25966

BEAS's accuracy and execution times for $N = 20$ and $N = 50$ clients.

Comparative Analysis

Framework	Comms†	Threat Model: Participants	Threat Model: Aggregator	Privacy Guarantees: Inference	Privacy Guarantees: Training	Security Guarantees: Data Poisoning	Security Guarantees: Model Poisoning	Security Guarantees: Byzantine Attack	Techniques Used: HE	SS+AE	TP	FE	DP	Blockchain	Features: Identity Privacy	Statistical Heterogeneity	Scalability	Asynchronous Updates	Dynamic Participants	Decentralized	Premature Convergence	Reward Mechanism	Open-Source
BinDaaS (Bhattacharya et al. 2019)	3 rounds	□	–	○	○	○	○	○	○	○	○	○	○	●	●	○	●	○	○	○	○	○	×
PiRATE (Zhou et al. 2020)	3 rounds	⊠	–	◐	◐	●	○	○	○	○	○	○	○	●	●	●	○	○	○	○	○	●	×
BAFFLE (Ramanan and Nakayama 2020)	3 rounds	■	–	●	●	●	○	○	○	○	○	○	○	●	●	●	○	○	○	○	○	●	×
Li et al. (Li et al. 2020)	3 rounds	⊠	–	○	○	●	○	○	○	○	○	○	○	●	●	○	●	○	○	○	○	●	×
LearningChain (Chen et al. 2018)	3 rounds	■	–	●	●	◐	◐	◐	○	○	○	○	●	●	●	●	○	○	○	○	○	●	×
Biscotti (Shayan et al. 2018)	3 rounds	■	–	●	●	●	◐	◐	○	○	○	○	●	●	●	●	●	○	○	○	●	●	×
POSEIDON (Sav et al. 2020)	2 rounds	■	⊠	●	●	○	○	○	●	○	○	○	○	○	●	●	○	○	○	●	○	○	×
Shokri et al. (Shokri and Shmatikov 2015)	1 round	□	□	○	○	○	○	○	○	○	○	○	○	○	◐	●	○	○	◐	○	○	○	×
PATE (Papernot et al. 2018)	1 round	□	□	○	○	○	○	○	○	○	○	○	○	○	○	◐	○	○	○	○	○	○	×
HybridAlpha (Xu et al. 2019)	1 round	■	⊠	●	◐	◐	○	○	○	○	○	●	●	○	●	●	○	○	●	○	○	○	×
BEAS(This Work)	1 round	■	–	●	●	●	●	●	○	○	○	○	●	●	●	●	●	●	●	●	●	●	✓

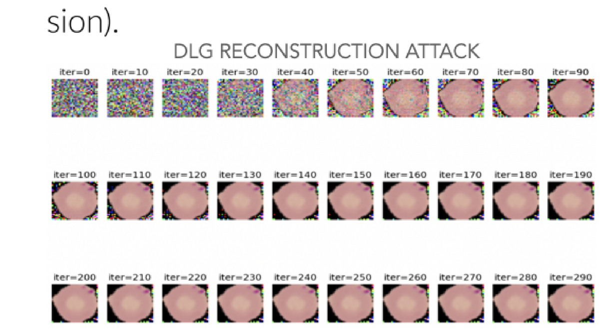
An overview of the BEAS protocol

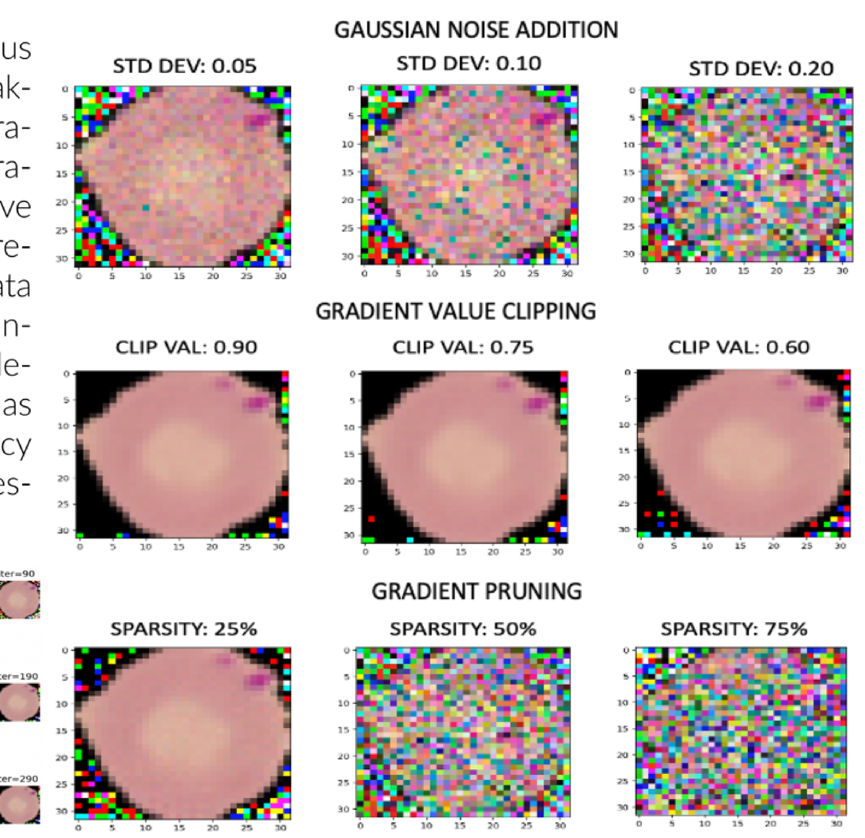


1. Clients create cryptographically anonymous identities using the MSP.
2. Genesis clients initiates the protocol by setting up a new channel, defining training parameters, and generating a genesis global block by training on their own data.
3. Participating clients request the previous global block to initialise a pre-training model, and update it by training on their own private datasets to generate new local gradients.
4. Client sends new local gradients to the EP, which creates a new local block and shares it with the ordering service.
5. Ordering service establishes consensus on the ordering of blocks, and commits them onto the ledger.
6. Once a threshold of local blocks is attained on the ledger, merge chain code is triggered to aggregate and create the new global block.
7. Steps 3 to 6 get repeated until desired accuracy is achieved, or ad-infinitum.

Privacy Guarantees

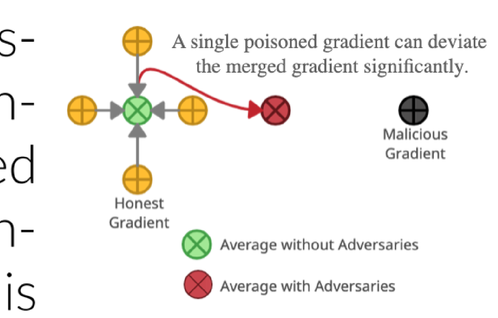
We implement and compare various DP techniques to prevent direct leakage of training data from shared gradients. Our experiments show gradient pruning (GP) is more effective than existing DP techniques: it prevents reconstruction of training data from shared model gradients with minimal impact on performance,and defends against model poisoning. GP has not been used in prior work for privacy (GP's primary use: gradient compression).





Security Analysis

We minimize risk of data poisoning using a combination of protocols to identify adversaries: (i) Multi-KRUM is used to guarantee resiliency from independent adversaries; and (ii) FoolsGold is used to identify Sybil groups.



Defense	Number of Adversaries: 0	1	5	10
NIL	96.16	96.02	82.88	57.20
MK	94.22	94.60	91.17	72.11
FG	95.63	82.11	87.50	85.72
MK + FG	94.16	90.26	87.24	83.66

BEAS accuracy with FoolsGold (FG) and Multi-KRUM (MK) under Label Flipping attack for different number of adversaries and (N = 20); Dataset: Malaria Cell Image.

Framework	Main Task Accuracy (%)			Backdoor Task Accuracy (%)		
Adversaries per Round	0	1	5	0	1	5
BEAS	96.16	95.81	96.08	11.06	28.20	61.85
BEAS + Noise (0.05)	85.84	84.66	82.10	09.76	19.44	49.16
BEAS + Clipping (0.80)	94.55	94.16	93.60	11.33	27.21	62.70
BEAS + Pruning (0.60)	92.95	92.67	92.88	10.20	22.46	43.88

Beas accuracy on main task and backdoor subtask with different differential privacy techniques under Pixel Pattern Backdoor Model Poisoning attack for different number of adversaries and (N = 20); Dataset: Malaria Cell Image.

Future Work

1. Improve resilience against membership-inference, property inference and linkability attacks.
2. Conduct tests using synthetic data for effective privacy preservation.
3. Deploy BEAS via open-source channels for different academic and industrial purposes to observe its working real-time.

Author Contact: harpreet.virk, arup.mondal_phd19, debayan.gupta, @ashoka.edu.in

Ashoka University, Rajiv Gandhi Education City, Sonipat, Haryana 131029