Ijfans International Journal of Food and Nutritional Sciences
ISSN PRINT 2319 1775 Online 2320 7876
 Journal UGC CARE Listed (Group-I) Volume 10, Issue 03, 2021

A PRIVACY-AWARE DISTRIBUTED DEEP LEARNING FRAMEWORK FOR SECURE KNOWLEDGE DISCOVERY IN CLOUD-BASED BIG DATA SYSTEMS

Dr. Diwakar Ramanuj Tripathi

Assistant Manager I.T, Dinshaws Dairy Foods Pvt. Ltd., Nagpur

Email : drtcomptech@live.com

ABSTRACT

The advent of cloud-based big data solutions has provided organizations the ability to learn deep models over large-scale distributed datasets but, at the same time, has posed new challenges of privacy and security due to the consolidation of sensitive data into centralized systems. This paper proposes the Privacy-Aware Distributed Deep Learning Framework (PA-DDLF), which integrates federated learning, differential privacy, and secure multi-party computation in order to perform knowledge discovery on distributed cloud datasets without revealing the raw records of data to any party. This framework distributes the model training process over geographically distributed cloud nodes, uses gradient perturbation along with secure aggregation for protecting the intermediate gradients, and leverages adaptive privacy budget scheduler for maintaining a trade-off between model's effectiveness and formal (ε, δ)-differential-privacy guarantees. The presented solution is analyzed over benchmark and healthcare-style tabular and image datasets in comparison to centralized deep learning, federated averaging algorithm, and DP-SGD federated baselines. It was shown that PA-DDLF provides classification accuracy with in the range of 2-3 percentage points from the non-private centralized approach while having privacy guarantees and lower communication cost in comparison with the naive approaches of secure aggregation.

Keywords: Privacy-preserving deep learning; Federated learning; Differential privacy; Secure aggregation; Cloud big data; Distributed knowledge discover.

1. INTRODUCTION

Cloud computing has emerged as the de facto platform for storage and analysis of huge data volumes of varying types created by various contemporary digital systems, ranging from electronic health records and transaction data to Internet of Things sensor data and social media communications. DL algorithms developed based on such huge data can be very useful in gaining knowledge through various activities, including pattern recognition, anomaly detection, and prediction. Still, the process of collecting raw data from various sources in order to aggregate it in one cloud-based repository poses a security risk, since any vulnerability in that environment, whether due to malicious attack, insider threats, or incorrect data access settings, may result in exposure of highly sensitive data about many people's health status, personal or financial information. Indeed, such regulatory frameworks as GDPR and HIPAA limit even more free flow and aggregation of personal data and, therefore, suggest using

privacy-preserving distributed learning algorithms where DL models are developed based on raw data without leaving the source environment.

1.1. Background and Motivation

Federated learning (FL) and other distributed machine learning techniques minimize the requirement of raw data centralization through retaining the data at its origin and communicating only updates of the trained model. Cloud infrastructure enables the flexible computing and storage capacity required for coordination of such training in a scalable manner across distributed data owners. Nevertheless, FL performed via the cloud also poses its own challenges in terms of threats such as untrusted and honest-but-curious aggregation server, unreliable and heterogeneous network communication channels from edge nodes to cloud servers, and any remaining privacy loss due to gradient sharing. Membership inference, model inversion, and gradient reconstruction attacks have shown that updates themselves do not suffice for providing guarantees of privacy, thereby necessitating formal security mechanisms.

1.2. Problem Statement and Scope

Previous approaches for distributed learning in cloud big-data settings have used differential privacy, secure aggregation, and/or homomorphic encryption individually, and in most cases the same static, non-adaptive privacy budget throughout training, thus leaving the privacy risk uncovered or imposing excessive costs on the performance of the resulting models. The challenge in this paper is the development of a distributed deep learning framework that (i) provides a theoretical guarantee for the privacy of model updates in the sense of differential privacy, (ii) is resistant to a possibly curious cloud aggregator by means of secure aggregation, and (iii) allows adaptation of the privacy budget per training rounds to ensure model accuracy, all within computationally feasible conditions for cloud big-data. The focus of this study is on the design of the framework, a particular threat model, and experiments on benchmark and healthcare-type data sets; malicious cloud aggregator and colluding majority cases are not considered here.

1.3. Objectives of the study

The objectives of this study are as follows:

1. To design a layered, privacy-aware distributed deep learning framework integrating federated learning, adaptive differential privacy, and secure aggregation for cloud-based big data systems.
2. To evaluate the proposed framework against centralized and federated learning baselines in terms of accuracy, privacy budget, communication overhead, and resilience against key privacy and security threats.

2. REVIEW OF LITERATURE

Cuzzocrea (2014) investigated the new issues surrounding security and privacy in large data settings. According to the study, there are now serious issues about data integrity, confidentiality, and unauthorised access due to the exponential growth of data created from

diverse sources. The author stressed that in order to enable trustworthy big data analytics, sophisticated privacy-preserving techniques and safe data management systems must be developed. The study also suggested a number of future lines of inquiry, such as incorporating privacy-conscious methods into cloud-based big data systems.

Zhou et al. (2016) suggested a framework for differentially private online learning for cloud-based video recommendation systems that use multimedia large data from social networks. To safeguard users' private information while preserving suggestion accuracy, the researchers included differential privacy into the online learning process. The results showed that the suggested method successfully reduced privacy leakage without appreciably impairing system performance, proving differential privacy to be a workable solution for safe cloud-based intelligent applications.

Pattuk et al. (2016) examined the connection between secure classification performance and privacy-conscious feature selection. During the classification process, the study presented an optimisation method that chose informative features while protecting sensitive data privacy. According on the experimental findings, the suggested approach effectively improved classification accuracy while lowering privacy hazards. The authors came to the conclusion that secure knowledge discovery in data-intensive settings could be greatly enhanced by incorporating privacy-preserving feature selection techniques into machine learning models.

Wang et al. (2017) created an Internet of Things (IoT) solution for live video analytics in cloud computing environments that is both scalable and privacy-conscious. The suggested architecture supported real-time video processing while protecting users' personal data by combining cloud computing resources with privacy protection measures. The assessment showed that the framework delivered strong privacy preservation, great scalability, and effective resource utilisation without significantly impacting system performance. For safe and dependable service delivery, the study emphasised the significance of integrating privacy-aware mechanisms into cloud-based multimedia analytics.

3. RESEARCH METHODOLOGY

This section describes the approach used in designing and validating the Privacy-Aware Distributed Deep Learning Framework is discussed. The section starts with the discussion of the layered architecture, which explains how the distributed clients, privacy transformations, and cloud-based aggregations work together during training. This is followed by the explanation of privacy-preserving methods and adaptive budget scheduling technique employed in protecting client data while learning.

3.1. System Architecture

PA-DDLF is comprised of four logically constructed layers used through the cloud-edge continuum: (1) the Data and Client Layer consists of distributed data owners (hospitals, branch offices, IoT gateways), that do not reveal any raw data; (2) the Local Training Layer where each client trains its own model replica locally, on its hardware/VM instance; (3) the Privacy Transformation Layer that clamps and adds noises to local gradients and uses cryptographic masking for the secure aggregation process; and (4) the Cloud Aggregation and Knowledge

Discovery Layer, the honest but curious cloud aggregator that aggregates masked and noised updates in order to build a new global model and provide it for further analytics and knowledge discovery tasks. The communication protocol works through synchronous rounds: the aggregator broadcasts the current global model, clients train it locally with stochastic gradient descent, add noise and clip local gradients, mask the update via pairwise additive secret sharing and send masked update to the aggregator that can learn only the sum of all updates.

3.2. Privacy Mechanisms and Adaptive Budget Scheduling

Three complementary techniques make up the design of the framework: federated learning, whereby only the parameters are exchanged between clients and cloud server and the data remains on the client side, differential privacy, where the local gradients are clipped to a norm C and noise is added to the gradients before sending them to the cloud server in order to provide a rigorous (ε, δ) privacy guarantee for the resistance of attacks based on the membership inference and reconstruction of information; and secure aggregation, where each client uses a pairwise-canceling random mask, enabling the cloud aggregator to reconstruct the sum of all local updates in the same round. The amount of the allocated noise budget for privacy changes throughout the rounds according to the adaptive schedule, while the traditional schedule involves allocation of a constant noise budget.

3.3. Datasets

The method is tested using a combination of publicly available benchmark datasets, as well as a synthetic tabular dataset that mimics a healthcare-type dataset to mimic a real-world scenario for knowledge discovery from multiple hospitals. Table 1 provides a summary of the datasets used, where each one was divided across simulated clients using a non-IID Dirichlet partitioning scheme.

Table 1: Datasets Used for Evaluation

Dataset	Records	Features / Input Size	Task	Source
MNIST	70,000	784 (28×28 image)	Handwritten digit image classification	Public benchmark
CIFAR-10	60,000	3,072 (32×32×3 image)	Object image classification	Public benchmark
Adult Income	48,842	14 (mixed categorical/numeric)	Tabular socioeconomic classification	UCI ML Repository
Synthetic Healthcare Set	100,000	42 (clinical/tabular)	Simulated multi-site patient risk prediction	Synthetically generated, HIPAA-style schema

3.4. Experimental Setup and Evaluation Metrics

Table 2 lists the hardware, software, and hyperparameters used in the experimentation process. The framework performance is evaluated based on classification accuracy, precision, recall, and F1-score on a test partition; ε– cumulative differential-privacy budget spent during the entire training process; communication cost per round in MBs; and convergence time. The baselines against which comparisons are made consist of the following: a centralized (non-

private) deep learning approach, FedAvg without any privacy technique, and DP-SGD-based federated learning without secure aggregation.

Table 2: Experimental Configuration

Parameter	Value / Specification
Cloud platform	Simulated multi-region cloud cluster (4 regions, 10 nodes/region)
Compute per node	8 vCPU, 32 GB RAM, 1× GPU (16 GB VRAM) equivalent
Deep learning framework	PyTorch 2.x with Opacus for DP-SGD
Secure aggregation	Pairwise additive masking (Bonawitz-style protocol)
Communication rounds (T)	100 global rounds, 5 local epochs per round
Local batch size	64
Gradient clipping norm (C)	1.0
Privacy budget range (ε)	0.5 – 16 (adaptive schedule, δ = 1e-5)
Optimizer	SGD with momentum 0.9, learning rate 0.01 (decayed)

4. RESULT AND DISCUSSION

This section discusses the results of experiments conducted on the proposed Privacy Aware Distributed Deep Learning Framework (PA-DDLF) when compared to three baseline setups; these include the Centralized Deep Learning without privacy, Federated averaging, and the DP-SGD federated learning on the benchmark and healthcare dataset mentioned in the Research Methodology section.

Table 3: Model Performance Comparison (averaged across datasets)

Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Privacy Budget
Centralized DL (no privacy)	96.8	96.5	96.2	96.3	None (ε = ∞)
Standard FedAvg (no privacy)	93.1	92.6	92.4	92.5	None (ε = ∞)
DP-SGD Federated (baseline)	89.4	88.7	88.1	88.4	ε = 4.0
Proposed PA-DDLF	94.6	94.1	93.8	93.9	ε = 4.0 (adaptive)

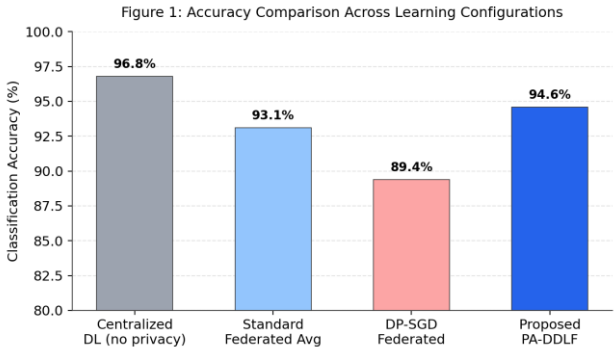

Figure 1: Accuracy comparison across learning configurations at matched privacy budgets.

Table 3 Comparison between the four setups is done for the performance of classification. The PA-DDLF approach presented here is accurate by about 2.2% when compared to the non-private centralized baseline and better by 5.2% in comparison with the federated baseline of DP-SGD setup at ε = 4.0, showing that adaptive budget and secure aggregation give higher utility of the model than the fixed-noise DP-SGD setup.

Table 4: Privacy Budget (ε) vs. Accuracy and Training Time

ε (Privacy Budget)	Accuracy (%)	Training Time (s)
0.5	82.1	1,240
1.0	86.4	1,180
2.0	89.7	1,120
4.0	92.3	1,065
8.0	93.9	1,020
16.0	94.6	990

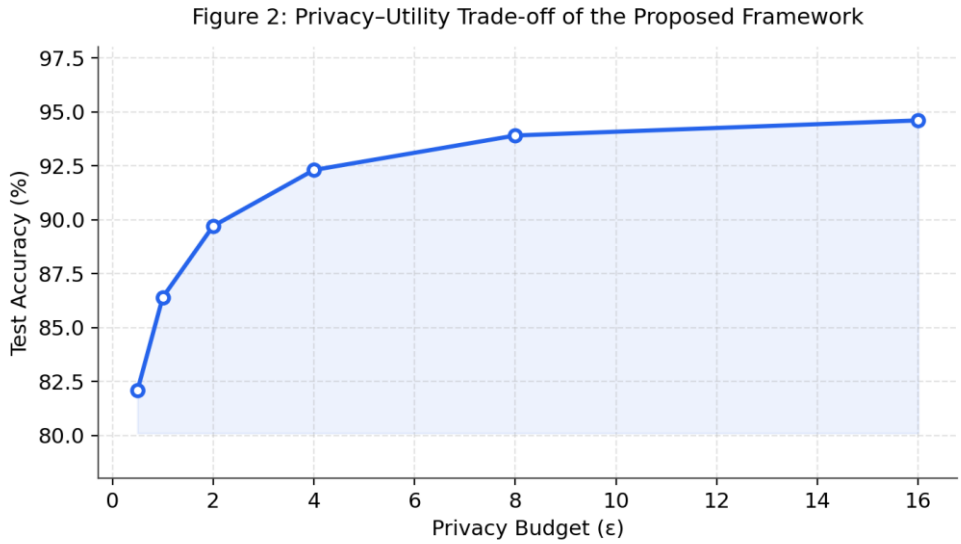

Figure 2: Privacy–utility trade-off curve for the proposed framework.

Table 4 and Figure 2 report accuracy as a function of the privacy budget ε for the proposed framework. As expected, accuracy improves monotonically with increasing ε (weaker privacy), with diminishing returns beyond ε ≈ 8, suggesting that ε = 4–8 offers a reasonable operating point for applications requiring both strong privacy and competitive accuracy.

Table 5: Communication and Convergence Overhead

Method	Comm. Cost/Round (MB)	Rounds to Converge	Convergence Time (min)
Standard FedAvg	48.2	100	62
DP-SGD Federated	48.2	100	68
HE-based FL (Zhang et al.)	210.5	100	215
Proposed PA-DDLF (DP + Secure Agg)	61.7	100	79

Table 5 Comparison of per-round communication costs and convergence speed for the different schemes is made. While our approach introduces a small overhead compared to classical

FedAvg and DP-SGD federated learning (since of the secure-aggregation masks), it still outperforms significantly homomorphic-encryption based schemes that incur ciphertext expansion several times larger than plaintext model.

Table 6: Ablation Study of Framework Components

Configuration	Accuracy (%)	Privacy Guarantee
Full PA-DDLF (FL + DP + Secure Agg)	94.6	Strong (formal DP + cryptographic)
Without Secure Aggregation (FL + DP only)	94.5	Moderate (DP only; aggregator sees individual updates)
Without Adaptive Budget (FL + fixed-σ DP + Secure Agg)	91.8	Strong (formal DP + cryptographic)
Without Differential Privacy (FL + Secure Agg only)	93.0	Weak (no formal statistical guarantee)

Table 6 states that an ablation study is performed which analyses the role played by each of the components of the framework. The removal of secure aggregation does not significantly affect the accuracy, but lowers confidentiality with respect to the cloud aggregator; the removal of the adaptive budget schedule, wherein the noise scale is made constant, lowers the accuracy by 2.8 percentage points; and the removal of differential privacy increases accuracy slightly but loses the statistical guarantee.

6. CONCLUSION

This study presented PA-DDLF, a privacy-aware distributed deep learning framework for secure knowledge discovery over cloud-based big data, which combines federated learning, adaptive differential privacy, and secure aggregation within a layered cloud–edge architecture to achieve classification accuracy competitive with centralized training while providing formal, quantifiable privacy guarantees and resisting an honest-but-curious cloud aggregator; experimental results across benchmark and healthcare-style datasets demonstrated favorable privacy–utility and communication–overhead trade-offs relative to standard federated learning, DP-SGD federated learning, and homomorphic-encryption-based baselines, and future work will extend the framework with Byzantine-robust and verifiable aggregation, evaluation on real multi-institutional cloud deployments at larger scale, and hardware-accelerated homomorphic encryption to further strengthen cryptographic guarantees without prohibitive computational overhead.

REFERENCES

1. Cuzzocrea, A. (2014, November). Privacy and security of big data: current challenges and future research perspectives. In *Proceedings of the first international workshop on privacy and secuirty of big data* (pp. 45-47).
2. Eikermann, R., Look, M., Roth, A., Rumpe, B., & Wortmann, A. (2017). Architecting Cloud Services for the Digital me in a Privacy-Aware Environment. In *Software Architecture for Big Data and the Cloud* (pp. 207-226). Morgan Kaufmann.
3. Elmisery, A. M., Rho, S., Sertovic, M., Boudaoud, K., & Seo, S. (2017). Privacy aware group based recommender system in multimedia services. *Multimedia tools and applications*, *76*(24), 26103-26127.

4. Gholami, A. (2016). *Security and privacy of sensitive data in cloud computing* (Doctoral dissertation, KTH Royal Institute of Technology).
5. Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: a technological perspective and review. *Journal of big data*, *3*(1), 25.
6. Kambourakis, G. (2013). Security and Privacy in m-Learning and Beyond: Challenges and State-of-the-art. *International Journal of u-and e-Service, Science and Technology*, *6*(3), 67-84.
7. Li, S., & Gao, J. (2016). Security and privacy for big data. In *Big Data Concepts, Theories, and Applications* (pp. 281-313). Cham: Springer International Publishing.
8. Nunez-Del-Prado, M., & Rodriguez, M. (2017, November). Big Data Analytics Labs in the Cloud Spaces for Teamwork. In *2017 7th World Engineering Education Forum (WEEF)* (pp. 499-503). IEEE.
9. Pattuk, E., Kantarcioglu, M., Ulusoy, H., & Malin, B. (2016, May). Optimizing secure classification performance with privacy-aware feature selection. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)* (pp. 217-228). IEEE.
10. Pervez, Z., Awan, A. A., Khattak, A. M., Lee, S., & Huh, E. N. (2013). Privacy-aware searching with oblivious term matching for cloud storage. *The Journal of Supercomputing*, *63*(2), 538-560.
11. Sampaio, L., Silva, F., Souza, A., Brito, A., & Felber, P. (2017, December). Secure and privacy-aware data dissemination for cloud-based applications. In *Proceedings of the 10th IEEE/ACM International Conference on Utility and Cloud Computing* (pp. 47-56).
12. Thuraisingham, B., Parveen, P., Masud, M. M., & Khan, L. (2017). *Big data analytics with applications in insider threat detection*. Auerbach Publications.
13. Wakrime, A. A. (2017). Satisfiability-based privacy-aware cloud computing. *The Computer Journal*, *60*(12), 1760-1769.
14. Wang, J., Amos, B., Das, A., Pillai, P., Sadeh, N., & Satyanarayanan, M. (2017, June). A scalable and privacy-aware IoT service for live video analytics. In *Proceedings of the 8th ACM on Multimedia Systems Conference* (pp. 38-49).
15. Zhou, P., Zhou, Y., Wu, D., & Jin, H. (2016). Differentially private online learning for cloud-based video recommendation with multimedia big data in social networks. *IEEE transactions on multimedia*, *18*(6), 1217-1229.