Joint Knowledge Distillation and Local Differential Privacy for Communication-Efficient Federated Learning in Heterogeneous Systems

Gad Gad*[1], Zubair Md Fadlullah*[2], Mostafa M. Fouda†‡[3], Mohamed I. Ibrahem§¶[4], and Nidal Nasser‖[5]
*Department of Computer Science, Western University, London, ON, Canada.
†Department of Electrical and Computer Engineering, Idaho State University, Pocatello, ID, USA.
‡Center for Advanced Energy Studies (CAES), Idaho Falls, ID, USA.
§School of Computer and Cyber Sciences, Augusta University, Augusta, GA, USA.
¶Department of Electrical Engineering, Faculty of Engineering at Shoubra, Benha University, Cairo, Egypt.
‖College of Engineering, Alfaisal University, Saudi Arabia.
Emails: [1]ggad@uwo.ca, [2]zfadlullah@ieee.org, [3]mfouda@ieee.org, [4]mibrahem@augusta.edu, [5]nnasser@alfaisal.edu.

***Abstract*—Federated Learning (FL) has emerged as a powerful approach to facilitate the construction of centralized models without compromising the data privacy of multiple participants. However, conventional FL methodologies do not address system heterogeneity where each participant needs to independently design its own model, a prevalent requirement in Internet of Things (IoT) applications due to the heterogeneous nature of tasks and data. Knowledge Distillation-based FL algorithms tackle this limitation by exchanging soft labels instead of model weights, thus giving each client the ability to independently design its local model architecture. While FL is inherently private, studies have indicated that exploiting gradients for a few iterations can reveal sensitive training data. To protect against privacy attacks, FL algorithms employ Differential Privacy (DP) to guarantee privacy protection, which can be applied using Local Differential Privacy (LDP). In this paper, we elaborate on preserving clients' training data privacy in KD (Knowledge Distillation)-based FL using DP, providing both privacy and flexibility. We provide theoretical analysis to extend the privacy guarantee to exchanged updates. Experimental analysis is performed utilizing Human Activity Recognition (HAR) datasets with different modalities. The results obtained demonstrate the capacity of KD-based FL to maintain a robust utility-privacy balance. Furthermore, for the same DP protection level, the utility of models trained on images was severely reduced across all FL algorithms. This suggests that the modality and complexity of a dataset are important factors for shaping the utility-privacy tradeoff of DP.**

***Index Terms*—Communication-efficiency, deep learning, heterogeneous federated learning, differential privacy, knowledge distillation, human activity recognition.**

I. Introduction

Training high quality deep learning models requires lots of labeled data. Moving and collecting data from edge devices, where they are born, to central servers poses privacy risks. Federated Learning (FL) [1]–[8] offers a decentralized approach enabling distributed Deep Learning (DL) model training without sharing users' data. Despite its inherent privacy-preserving properties, FL can still be vulnerable to privacy attacks, as studies have shown that exploiting gradients for a few iterations can reveal sensitive training data [9]. To address this issue, privacy-preserving techniques, such as Differential Privacy (DP) [10] and the Private Aggregation of Teacher Ensembles (PATE) [11] were developed.

Another shortcoming of model-based (FedAvg) FL arises in heterogeneous communication systems in which each participant needs to independently designs their own model. This requirement is particularly prevalent in the Internet of Things (IoT) applications, due to the heterogeneous nature of tasks and data, as well as the variability of resources available on each user-device.

Knowledge Distillation (KD) [12], [13], which was originally developed to transform knowledge from a trained large model (called the teacher) to a smaller to-be-trained model (called the student) in the central setting, has been repurposed in our work to the federated setting. In addition to the significant communication overhead advantage of KD-based FL algorithms compared to model-based FL algorithms, the former also enables clients to independently design their local model architecture.

In this paper, we elaborate on applying DP to KD-based FL algorithms as depicted in Fig. 1. We commence by introducing KD-based FL as a communication-efficient FL algorithm that gives each client flexibility to independently design its local model. We then discuss the Differentially Private-Stochastic Gradient Descent (DP-SGD) algorithm, a mechanism to construct DP training pipelines for DL models to enforce a certain privacy protection level. Different algorithms were proposed to apply DP in the federated setting. Among them, we contrast Local Differential Privacy (LDP) with Central Differential Privacy (CDP). The LDP approach is adopted as it assumes a realistic FL threat model in which the server is trusted to comply with the FL protocol but is not trusted to guarantee clients' privacy. Figure 1 provides an overview of the proposed KD-based FL approach with DP.

The main contributions of our paper are summarized as follows.

This research was partly supported by the New Frontiers in Research Fund (NFRF) grant number: NFRFE-2019-00313, and the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant no. RGPIN-2020-06260.

- We propose an FL algorithm that seamlessly combines KD and DP. While KD-based FL algorithms are communication-efficient algorithms that also enable clients to independently design their local model, DP is added to provide privacy protection.
- We conduct theoretical analysis to extend the privacy guarantee associated with Local Differential Privacy (LDP) to the exchanged soft labels.
- Extensive experiments on Human Activity Recognition (HAR) datasets of different modalities were carried out. Our findings suggest that the characteristics of a dataset, including its modality and complexity, are essential in defining the utility-privacy tradeoff.

The remainder of the paper is structured as follows. Section II presents the recent research work appearing in the relevant domains. Our considered problem is described in section III. We describe our proposed methods in section IV. In section V, we evaluate the preformance of our proposed methodology through extensive experiments and discuss the obtained results. Finally, section VI concludes the paper.

II. Related Work

Federated learning (FL) is a decentralized, privacy-preserving learning paradigm. In FL, distributed devices collaborate to train their local models on private data by sharing model updates [1], [5]. System heterogeneity and communication overhead are two major challenges facing the deployment of FL applications. Efforts have been directed toward improving communication-efficiency via asynchronous weight updates [14] and gradient compression [15], though the latter approach may result in performance degradation with high compression ratios.

Deep learning models learn by memorizing patterns and data, therefore it necessary to ensure the privacy of training data [16]. Differential privacy [10], [17] is a mathematical framework for privacy analysis that is used to design private training pipelines for deep learning models to provide formal privacy protection to training data samples against privacy attacks. In particular, differential privacy has been shown to effectively protect against Membership Inference Attacks (MIA) in which an adversary is interested to know whether a given sample was used to train a neural network [9]. Recent works have also demonstrated that by exploiting implicit memorization, sensitive data can be revealed not only from the model parameters but also from the model output. For example, Fredrikson et al [18] retrieved training samples (individual faces) by exploiting the output probabilities of a computer-vision classifier. DP ensures the privacy of individual samples in the training data by first bounding each sample's contribution to the learning process by modifying the Stochastic Gradient Descent (SGD) algorithm by clipping per-sample gradients and injecting calibrated noise during model training.

Knowledge Distillation was first proposed for the central training setting [12], [19] to compress the knowledge obtained by a large model (called the teacher) to a smaller model (called the student). In this approach, the teacher is first trained using a labeled dataset, then the smaller model is trained in an alternating fashion on the labeled dataset and to minimize the distance between its soft labels and the teacher's soft labels. Modifications to this technique were introduced to adapt KD for the Federated Learning (FL) scenario [20]–[22]. The authors in [20] employed a proxy dataset D_p, which is accessible to all clients for soft labels. Each client calculates his soft labels SL using the locally trained model on the shared dataset. Subsequently, the local soft labels are transmitted to a server for aggregation into global soft labels, which are then returned to the clients for training on the labeled dataset (D_p, SL^r). To improve the efficiency of knowledge distillation, Gad et al. [21] proposed FedAKD which employs Mixup augmentation [23] to generate a new version of the public dataset D_p each round synchronized across all clients using two server-controlled parameters.

KD-based FL methods address both system heterogeneity and communication overhead challenges as explored in our coauthors' earlier research work [24], [25]. Since the shared soft labels transfer knowledge across the network in KD-based FL algorithms, adversaries become interested in extracting private training data from the exchanged updates. While many previous studies explored applying DP in model-based FL algorithms [26]–[28], KD-based FL algorithms received less interest [29], [30]. The authors in [30] propose two algorithms: FedMD-NFDP which offers a noise-free privacy guarantee and FedMD-LDP which injects noise to provide a privacy guarantee to protect the shared model predictions in KD-based FL. FedMD-NFDP calculates the soft labels using a subset of the public dataset randomly sampled each round. sampling with replacement was found to be consistently more private than sampling without replacement [30]. In this work, we apply LDP to the training data, instead of the soft labels, and extend the privacy guarantee according to the post-processing immunity property.

III. Problem Description

Federated Learning can be mathematically formulated as follows. Let there be m participants in the learning process, denoted by P_1, P_2, ..., P_m, each owning a small labeled dataset $D_k = (x_{ki}, y_i)_{i=1}^{N_k}$, where k represents the participant index, and N_k represents the number of data points owned by participant k. These datasets may or may not be drawn from the same distribution.

While FL frameworks recently have a number of advancements with elegant and fast-converging algorithms approximating centralized trained model performance, heterogeneity, and communication overhead remain a unique challenge for distributed model training, especially in IoT environments. In such heterogeneous environments, communication-efficiency is much desired. At the same time, the native embedding of privacy remains at the heart of FL framework design to ensure that the privacy of the user data is preserved. In other words, communication-efficiency and privacy-embedding are two inter-bound problems, and intertwining them is not considered in the vanilla FL algorithms. Providing emphasis on

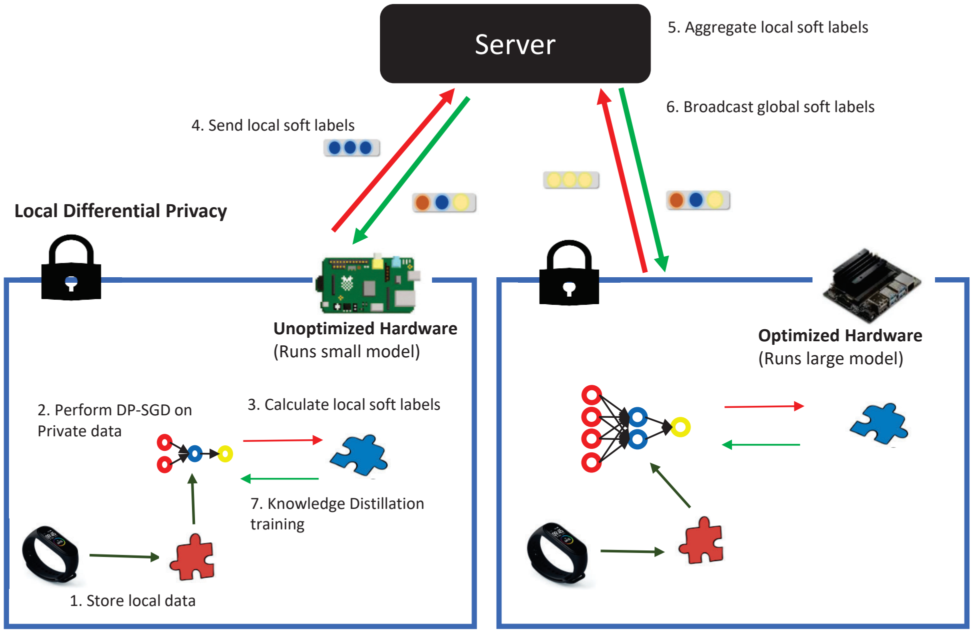
Server
5. Aggregate local soft labels
6. Broadcast global soft labels
4. Send local soft labels
Local Differential Privacy
Unoptimized Hardware
(Runs small model)
2. Perform DP-SGD on Private data
3. Calculate local soft labels
7. Knowledge Distillation training
1. Store local data
Optimized Hardware
(Runs large model)

Fig. 1. Considered Federated Learning (FL) system model with model heterogeneity phenomenon, and our treatment of this issue via joint fusion of Knowledge Distillation (KD) and Differential Privacy (DP).

one over the other can be regarded as a multi-objective decision problem. In this paper, we address this problem and design appropriate methodologies in the following section to incorporate both communication-efficiency and privacy-embedding in a seamless fashion.

IV. Envisioned Solution: Joint Knowledge Distillation and Local Differential Privacy for Communication-Efficient FL

In this section, we present our proposed methodologies that systematically combine Knowledge Distillation (KD), and Differential Privacy (DP), into Federated Learning framework. The rationale behind their joint incorporation is also theoretically analyzed in the section. Figure 1 depicts an overview of the proposed privacy-preserving FL approach.

A. *Knowledge Distillation-based Federated Learning*

In model-based FL [1], a central model architecture f with initial weights θ^0 is enforced by the server. During each round, the clients obtain a copy of the global weights θ_k^r, where r denotes the current round and k represents the client number. Clients train on local private data in parallel for few epochs before uploading their trained weights $\theta_k^{r,'}$ to be aggregated as follows,

$$\theta^{r+1} \leftarrow \sum_{i=1}^{N_k} \frac{P_i^r \cdot \theta_k^r}{\sum_{k=0}^{N_k} P_k^r}, \tag{1}$$

where θ^{r+1} represents the weights aggregated at round r and to be downloaded at round $r+1$. P_k^r denotes the test accuracy of the k-th client in round r used to weight clients' contributions by their respective performances.

Unlike FedaAvg [1], which imposes a server-controlled architecture to enable model aggregation, our envisioned KD-based FL addresses the system heterogeneity concern (which arises due to the variability of resources) by allowing clients to customize their own local model architectures. To achieve this, the clients in KD-based FL are assumed to exchange soft labels SL instead of model weights.

Next, we describe the KD-based FL approach, which is adopted with modifications in the studies described earlier [20]–[22]. Each global round in KD-based FL can be divided into four phases:

1) Local training phase (L): Locally designed models are trained on local data:

$$\theta_k^{r,'} \leftarrow \theta_k^r - \eta \cdot \frac{1}{|D_k|} \sum_{(x,y)\in D_k} \nabla \mathcal{L}_{CCE}(\theta f_k, f_k(x), y) \tag{2}$$

2) Upload Soft Labels phase (USL): Each client calculates his local soft labels $SL_k = f_k^*(D_p)$. and uploads SL_k to be aggregated by the server as
$f_k^*(x)$ is a model obtained from f_k by removal of the final Softmax layer [20].
3) Download Soft Labels (DSL): Global soft labels are downloaded to clients.

$$SL^r \leftarrow \sum_{i=1}^{N_k} \frac{P_i^r \cdot SL_k^r}{\sum_{k=0}^{N_k} P_k^r} \tag{3}$$

4) Knowledge Distillation Learning phase (KDL): Distance functions, such as Mean Squared Error (MSE) or Kullback-Leibler divergence loss (KD), are employed to train f_i on the public unlabeled dataset, utilizing the global soft labels.

$$\theta_k^{r+1} \leftarrow \theta_k^{r,'} - \eta \cdot \frac{1}{|D_p|} \sum_{(x,y)\in(D_p, SL^r)} \nabla \mathcal{L}_{KD}(\theta_k^{r,'}, f_k^*(x), y) \tag{4}$$

Here, $\theta_k^{r,'}$ and θ_k^{r+1} denote the model weights following local dataset training and model weights post knowledge distillation training, respectively. θ_k^{r+1} represents the weights of the k-th client at round $r+1$.

B. Preserving Privacy in Federated Networks: Local Differential Privacy (LDP)

Next, we investigate how to natively incorporate privacy-preservation in our considered FL framework through Local DP. To apply DP in the context of FL to protect clients' private data, two widely used approaches are [28]:

1) Central Differential Privacy (CDP): CDP provides participant-level protection. This setting assumes an honest server that is trusted by clients to inject noise.
2) Local Differential Privacy (LDP): LDP provides sample-level protection. Each client is responsible for applying DP-SGD in local training to produce a (ϵ, δ)-DP model that guarantees protection to each sample with probability bounded by ϵ.

In our work, we conisder the LDP by deriving inspiration from a theoretical work in [17] to protect the privacy to protect training data called Differentially Private Stochastic Gradient Descent (DP-SGD). DP-SGD achieves sample-level protection by clipping the gradients generated in the backpropagation cycle of the SGD algorithm. By clipping the gradient of each sample, the contribution of each sample to the output of the model is bounded. After that, a batch of gradients is averaged and noise is added according to the Gaussian Mechanism (GM).

C. Extending Privacy Guarantee to Soft Labels

Next, we aim to extend the privacy guarantee to soft labels while achieving the desired communication-efficiency in the considered FL system. Privacy attacks in model-based FL algorithms usually attempt to exploit exchanged gradients to reconstruct data [18] or to infer whether a given data point exists in the training data [9].

In KD-based FL algorithms [20], [21], [30], instead of weights, clients share knowledge in the form of soft labels. Sensitive data can be revealed not only from the model parameters but also from model predictions. For instance, Fredrikson et al. [18] retrieved training samples (individual faces) by exploiting the output probabilities of a computer-vision classifier.

In the case of KD-FL, outsider adversaries become interested in extracting private data from the shared model predictions (soft labels). As the shared soft labels are employed for knowledge distillation training of clients, it is difficult to maintain an acceptable utility-privacy balance with noise injection mechanisms that add noise to these predictions [31]. An alternative is to implement LDP to ensure that local model training satisfies a (ϵ, δ)-DP, and then employ the DP property of post-processing immunity to extend the protection of the differentially private trained model to its output predictions. The post-processing immunity property of DP states that the output a (ϵ, δ)-DP mechanism is also (ϵ, δ)-DP private.

We denote $M : N^{|X|} \to R$ as a randomized algorithm that satisfies (ϵ, δ)-differentially private. Additionally, $f : R \to R_0$ represents another function. Then, $f \circ M : N^{|X|} \to R_0$ also satisfies (ϵ, δ)-DP.

Proof. For a deterministic function $f : R \to R_0$.

For any neighboring databases x, y with $||x - y||_1 \leq 1$, and event $S \subseteq R_0$. Let $T = \{r \in R : f(r) \in S\}$. We then have:

$$\begin{aligned}\Pr[f(M(x)) \in S] &= \Pr[M(x) \in T] \\ &\leq \exp(\epsilon)\Pr[M(y) \in T] + \delta \\ &= \exp(\epsilon)\Pr[f(M(y)) \in S] + \delta\end{aligned}$$

□

This result demonstrates that the probability $\Pr[f(M(x)) \in S]$ is bounded by

$$\exp(\epsilon)\Pr[f(M(y)) \in S] + \delta. \tag{5}$$

This shows that applying the deterministic function f to the output of the (ϵ, δ)-differentially private algorithm M maintains the (ϵ, δ)-differential privacy guarantees. The result follows for randomized mapping as any randomized mapping can be decomposed into a convex combination of deterministic functions.

The privacy guarantee, which is attained by LDP where noise is injected into the gradients by each client during local training is inherited by any other algorithm applied to the output of these LDP-trained models, as shown in Fig. 2.

V. Performance Evaluation

In this section, by conducting extensive experiments based on multiple datasets across heterogeneous systems, we evaluate the performance of our proposed solution.

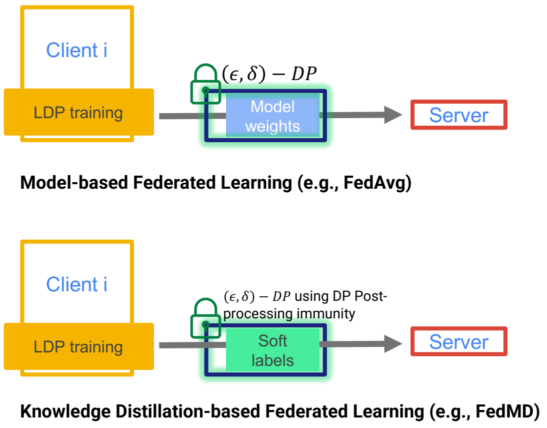
Client i
LDP training
$(\epsilon, \delta) - DP$
Model weights
Server
Model-based Federated Learning (e.g., FedAvg)
Client i
LDP training
$(\epsilon, \delta) - DP$ using DP Post-processing immunity
Soft labels
Server
Knowledge Distillation-based Federated Learning (e.g., FedMD)

Fig. 2. The privacy guarantee applied by a client with DP-SGD is inherited by the soft labels generated by that client according to the Post-Processing Immunity property.

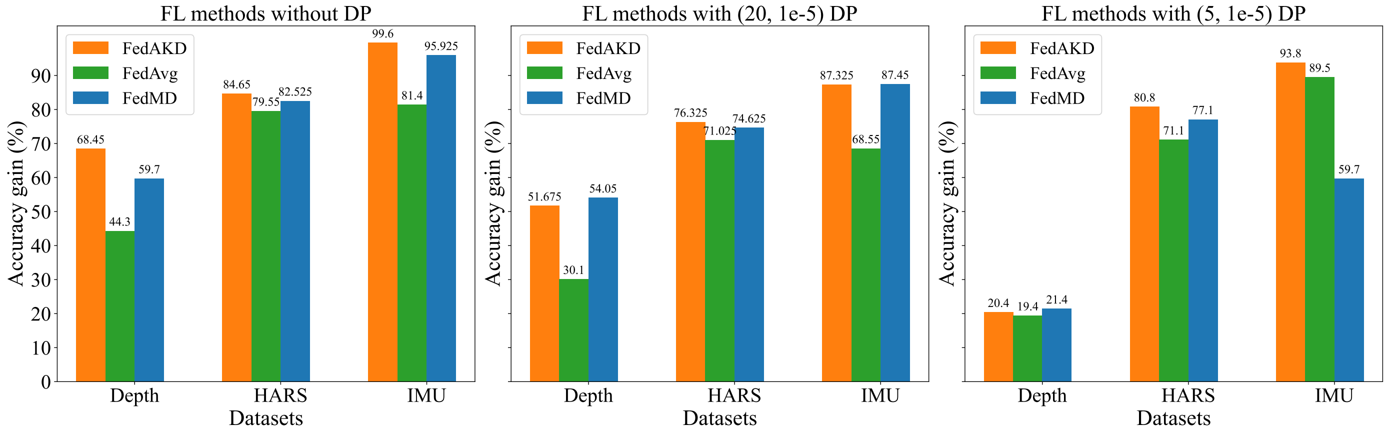

Fig. 3. Test accuracy of the FL methods on Depth, HARS, and IMU datasets with different differential privacy protection levels.

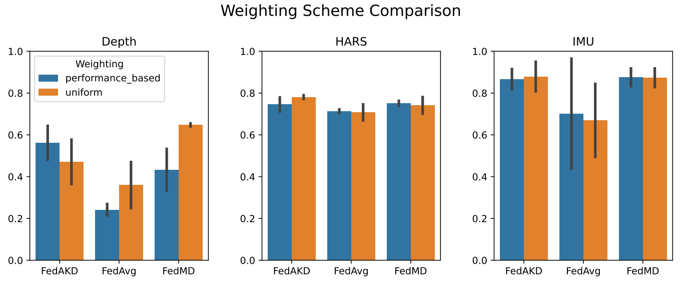

Fig. 4. Test accuracies of the FL methods on Depth, HARS, and IMU datasets. Bars are grouped based on the weighting scheme employed: uniform vs performance-based weighting while applying DP.

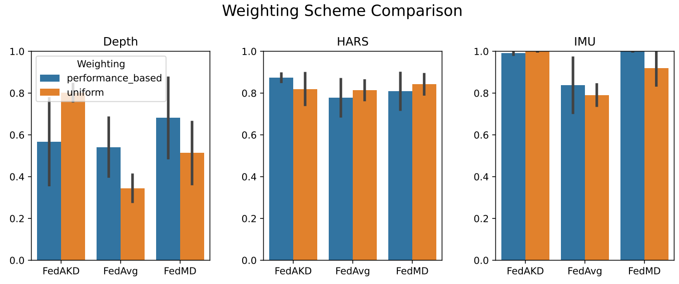

Fig. 5. Test accuracies of the FL methods on Depth, HARS, and IMU datasets. Bars are grouped based on the weighting scheme employed: uniform vs performance-based weighting while not applying DP.

Introducing noise to gradients in accordance with DP-SGD safeguards users' data; however, this results in a reduction in utility due to the clipping of gradients followed by the addition of noise to these clipped gradients. To assess the ability of various FL algorithms, such as FedAvg [1], FedMD [20], and FedAKD [21], to preserve utility across different ϵ values, we carried out experiments to examine the effects of incorporating noise under distinct protection levels on models trained with data of diverse modalities. We consider three DP protection levels: No DP, (20, 1e-5)-DP, and (5, 1e-5)-DP. We employed three Human Activity Recognition (HAR) datasets in our FL experiments: Depth [32], HARS [33], and IMU [32]. The first dataset is an image dataset while the other two datasets are tabular datasets.

Fig. 3 demonstrates three bar plots. From left to right, the accuracies obtained under the three privacy protection levels: No DP, (20, 1e-5)-DP, and (5, 1e-5)-DP are shown, respectively. For the Depth dataset in the left-most bar plot, we can see Knowledge Distillation-based FL methods, FedAKD/FedMD outperform FedAvg by over 15%. Applying differential privacy resulted in significant accuracy loss in all algorithms until they reach an accuracy of 20%. For the other two datasets, HARS and IMU, balancing utility-privacy is easier than that in the Depth dataset, under both protection levels ($\epsilon = 20$ and $\epsilon = 5$). The accuracy was less severe for HARS and IMU datasets than the accuracy drop observed in the Depth dataset for all FL algorithms. One of the reasons for this is the large dimension size and complexity of the Depth dataset relative to the two other datasets.

Figs. 4 and 5 demonstrate the test accuracies obtained by model-agnostic (FedMD/FedAKD) and model-based (FedAvg) FL algorithms using different weighting schemes for clients updates, respectively. We consider two weighting schemes, the uniform weighting scheme where clients contributions are weighted uniformly, and the performance-based weighting scheme where clients' contributions are weighted according to their test accuracy each round as in eqs. 1 and 3. Fig. 4 reports the accuracy while applying LDP and Fig. 5 shows the accuracy without applying DP. We can notice that the blue and orange bars that correspond to performance-based and uniform weighting, respectively, experienced close accuracy degradation, suggesting that varying the weighting scheme does not have any notable impact when it comes to balancing the utility-privacy tradeoff under DP.

VI. Conclusion

This paper presented a method for preserving clients' training data privacy in Knowledge Distillation (KD)-based Federated Learning (FL) using Differential Privacy (DP). The proposed approach has been designed in such a manner so that it allows clients to independently design their local model architecture while maintaining an efficient utility-privacy balance. Theoretical analysis was carried out to extend the Differential Privacy (DP) protection level obtained by LDP to the shared soft labels employing the Post-Processing Immunity property of DP. We also conducted experiments to explore the utility-privacy tradeoff of different FL algorithms, such as FedAvg, FedMD, and FedAKD by applying different DP protection levels using Local Differential Privacy (LDP) on Human Activity Recognition (HAR) datasets covering image and tabular modalities. Experimental results revealed that applying differential privacy leads to utility degradation across all algorithms. However, we found that the highest accuracy degradation affects models trained on image datasets. On the other hand, balancing utility and privacy was found to be easier for HARS and IMU datasets under the same protection levels. Thus, in summary, the proposed KD-based FL method can be regarded to have the ability to offer a communication-efficient alternative to the conventional FedAvg algorithm. Our proposal was also found to enable client devices to independently design their local models, and have the capacity to maintain a robust utility-privacy balance across different datasets and protection levels. Moreover, we found that the modality and complexity of a dataset play a crucial role in shaping the utility-privacy tradeoff.

References

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," *arXiv preprint arXiv:1602.05629*, 2016.

[2] Z. M. Fadlullah and N. Kato, "HCP: Heterogeneous computing platform for federated learning based collaborative content caching towards 6G networks," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 1, pp. 112–123, 2022.

[3] Q. Yang, L. Fan, and H. Yu, *Federated Learning: Privacy and Incentive*. Springer Nature, 2020.

[4] Z. M. Fadlullah and N. Kato, "On smart IoT remote sensing over integrated terrestrial-aerial-space networks: An asynchronous federated learning approach," *IEEE Network*, vol. 35, no. 5, pp. 129–135, 2021.

[5] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[6] Y. Gupta, Z. M. Fadlullah, and M. M. Fouda, "Toward asynchronously weight updating federated learning for AI-on-edge IoT systems," in *2022 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, 2022, pp. 358–364.

[7] S. Sakib, M. M. Fouda, Z. Md Fadlullah, and N. Nasser, "On COVID-19 prediction using asynchronous federated learning-based agile radiograph screening booths," in *ICC 2021 - IEEE International Conference on Communications*, 2021.

[8] M. M. Badr, M. I. Ibrahem, M. Mahmoud, W. Alasmary, M. M. Fouda, K. H. Almotairi, and Z. M. Fadlullah, "Privacy-preserving federated-learning-based net-energy forecasting," in *SoutheastCon 2022*.

[9] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*, 2017.

[10] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[11] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *arXiv preprint arXiv:1610.05755*, 2016.

[12] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819.

[13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[14] A. Asad, M. M. Fouda, Z. M. Fadlullah, M. I. Ibrahem, and N. Nasser, "Moreau envelopes-based personalized asynchronous federated learning: Improving practicality in network edge intelligence," in *GLOBECOM 2023 - 2023 IEEE Global Communications Conference*, 2023.

[15] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep gradient compression: Reducing the communication bandwidth for distributed training," *arXiv preprint arXiv:1712.01887*, 2017.

[16] D. Arpit *et al.*, "A closer look at memorization in deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 2017, p. 233–242.

[17] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pp. 308–318.

[18] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 2015, pp. 1322–1333.

[19] T. Kim, J. Oh, N. Kim, S. Cho, and S.-Y. Yun, "Comparing kullback-leibler divergence and mean squared error loss in knowledge distillation," *arXiv preprint arXiv:2105.08919*, 2021.

[20] D. Li and J. Wang, "FedMD: Heterogenous federated learning via model distillation," *arXiv preprint arXiv:1910.03581*, 2019.

[21] G. Gad and Z. Fadlullah, "Federated learning via augmented knowledge distillation for heterogenous deep human activity recognition systems," *Sensors*, vol. 23, no. 1, article no. 6, 2022.

[22] C. Wu, F. Wu, L. Lyu, Y. Huang, and X. Xie, "Communication-efficient federated learning via knowledge distillation," *Nature communications*, vol. 13, no. 1, p. 2032, 2022.

[23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[24] G. Gad, Z. M. Fadlullah, K. Rabie, and M. M. Fouda, "Communication-efficient privacy-preserving federated learning via knowledge distillation for human activity recognition systems," in *ICC 2023 - IEEE International Conference on Communications*, 2023.

[25] G. Gad, A. Farrag, Z. M. Fadlullah, and M. M. Fouda, "Communication-efficient federated learning in drone-assisted iot networks: Path planning and enhanced knowledge distillation techniques," in *2023 IEEE Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2023.

[26] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek, and H. Vincent Poor, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 3454–3469, 2020.

[27] S. Truex, L. Liu, K.-H. Chow, M. E. Gursoy, and W. Wei, "LDP-Fed: Federated learning with local differential privacy," in *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, 2020, pp. 61–66.

[28] M. Naseri, J. Hayes, and E. De Cristofaro, "Local and central differential privacy for robustness and privacy in federated learning," *arXiv preprint arXiv:2009.03561*, 2020.

[29] D. Gao and C. Zhuo, "Private knowledge transfer via model distillation with generative adversarial networks," *arXiv preprint arXiv:2004.04631*, 2020.

[30] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *International Conference on Machine Learning*, 2021, pp. 12 878–12 889.

[31] L. Sun and L. Lyu, "Federated model distillation with noise-free differential privacy," *arXiv preprint arXiv:2009.05537*, 2020.

[32] X. Ouyang, Z. Xie, J. Zhou, G. Xing, and J. Huang, "ClusterFL: A clustering-based federated learning system for human activity recognition," *ACM Transactions on Sensor Networks (TOSN)*, vol. 19, no. 1, article no. 17, 2022.

[33] D. Anguita *et al.*, "A public domain dataset for human activity recognition using smartphones," in *2013 European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2013, pp. 437–442.