Mitigating Membership Inference Attacks in Federated Learning

Fatima Elhattab, Sara Bouchenak

Liris,INSA, Lyon - France
firstname.lastname@insa-lyon.fr

Abstract
Federated Learning (FL) is a machine learning technique that allows multiple data owners to collaborate in training a model without sharing their training data. However, FL systems are still vulnerable to privacy attacks, where malicious participants can infer information about other participants' data by using the exchanged parameters. Membership inference attacks are a type of privacy attack that allows an attacker to infer whether a data sample was used by a target participant to train its own local model, thereby threatening data privacy. To address this issue, the paper proposes *PASTEL*, a novel FL privacy-preserving mechanism based on internal generalization gap minimization. The experimental evaluation of *PASTEL* shows that it reduces the membership inference attack success rate closely to 50% (best-case scenario) with a negligible impact on local models' utility.

Keywords : Distributed Machine Learning, Federated Learning, Membership Inference Attack, Privacy.

1. Background and Related Work

1.1. Federated Learning

Federated Learning (FL) is a machine learning framework (ML) that ensures privacy by distributing learning tasks to where the data is generated, such as on edge or mobile devices [24]. FL allows multiple workers to jointly train an ML model while keeping their raw data on their devices and transferring only the model parameters to the FL server. The workers train the models on their local data and send the model updates to a FL server, which aggregates the received updates using aggregation algorithms such as FedAvg [14], Federated Stochastic Variance Reduced Gradient [10], FedPer [1], or FedMa [22]. The FL server relies on secure aggregation [2], a secure multi-party computation protocol that ensures the aggregated model is produced in a privacy-preserving manner, preventing servers from examining clients' models.

1.2. Privacy Threats in Federated Learning for Edge Computing

Federated learning, which is designed to preserve privacy, is still vulnerable to privacy attacks, including membership inference attacks [20] that determine if a specific data record was used to train a model, and property inference attacks that extract dataset properties irrelevant to the training task. Model inversion or attribute inversion attacks [7, 6] are another type of privacy attack that attempt to recover sensitive features or full data samples based on partial knowledge of some features and given output labels.

In this work, we are interested in the membership inference attack (MIA), a privacy attack that aims to determine if a specific data record is used in the training of the target model. The authors in [20] introduce a black-box attack that relies on the output class probability distribution of the model. In this scenario, the attacker trains one or several shadow models to generate data, which is then used to train multiple attack models (one for each class). Using confidence scores as inputs, these attack models output the membership status of the given record. An extension of [20] attack is proposed in [19] which is based on a single shadow model and relaxes the assumption that the shadow model is constructed the same way as the target model.

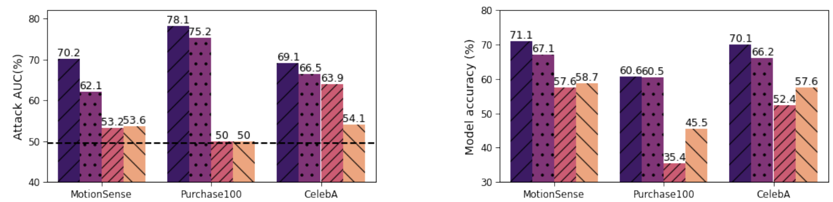
1.3. Related Work on Privacy Preservation Against Membership Inference in Edge Federated Learning

Recent work on mitigating membership inference attacks in Federated Learning (FL) can be divided into three categories : perturbation, cryptographic, and gradient compressibility methods. Perturbation methods such as Differential Privacy (DP) inject random noise to protect against information leakage, and can be divided into local and central DP. LDP and CDP have been found to reduce the performance of membership inference attacks in FL, as shown in [17, 21]. On the other hand, cryptographic methods such as Homomorphic Encryption and Secure Multi-Party Computation can guarantee privacy and accuracy, but they come with a significant cost on computation and memory [12]. HybridAlpha [3] is a privacy-preserving FL approach that employs an SMC protocol based on functional encryption, and shows significant improvements in terms of training time and data transfer volume. Gradient compressibility and sparsity have been shown to reduce the sources of information for privacy inference attacks [12, 5]. However, these approaches slightly degrade the performance of the global model, as reported in [26]. Therefore, there is a need to find the best trade-off between privacy and utility in FL [17, 9].

1.4. On the Difficulty of Mitigating Membership Inference Attacks in Edge Federated Learning

Differential privacy has been widely used as a framework for privacy-preserving machine learning, providing statistical guarantees against the information an adversary can infer through the output of a randomized algorithm. In the following, we describe a case that illustrates the problem of privacy leakage and the limitations of State-of-the-art systems in FL-based computing systems, in healthcare applications, computer vision and e-commerce applications . We focus on white-box inference membership inference attack proposed by [18] for classification tasks. To evaluate privacy leakage in healthcare applications, we used MotionSense dataset [13], which includes time-series data generated by accelerometer and gyroscope sensors (attitude, gravity, user acceleration, and rotation rate), For MotionSense we consider the classification task of determining the patient activity, for Purchase100 we train a classifier for determining the client type based on his purchases, and finally for CelebA the task consists in face attributes classification. We evaluate the attack with 3 differential privacy techniques namely WDP, LDP and CDP. The results for privacy leakage and model utility are presented in Figure 1.

Weak diffential privacy fails to protect againt membership inference attacks since it results in very large values of ϵ, as it adds noise at every round ignoring the noise added in previous rounds. More specifically, in DP, the concept of composability ensures that the joint distribution of the outputs of differentially private mechanisms satisfies DP [15]. Therefore, if we assume that, at every round, the server applies an ϵ-differentially private mechanism on participants' updates, then this weak DP mechanism results in spending $r x \epsilon$ privacy budget after r number



(a) Privacy attack success (b) Model accuracy

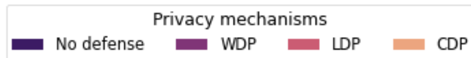


FIGURE 1 – Privacy leakage from edge computing image analysis – Impact on FL clients' models protected with existing FL privacy preserving mechanisms. The dashed line in first plot indicates the optimal privacy value.

of rounds. This yields larger values of ϵ, and thus significantly less privacy for participants. Concerning central differential privacy and local differential privacy, they provide to be efficient in mitigating membership inference attacks as shown in Figure 1, the attack AUC is reduced significantly. However, CDP and LDP come at a cost of sacrificing the performance of the model (it decreases from 60.6% to 35.4% with Purchase100 with LDP). Moreover, the main issue of differential privacy is the computation time, as more calculations are needed to add noise and other privacy-preserving operations.

2. Design Principle of *PASTEL*

We present *PASTEL* (*P*rIv*A*cy pre*S*erving federa*TE*d *L*earning), a local side privacy protection scheme that counters membership inference attacks in FL systems, without breaking secure aggregation guarantees, nor deteriorating the performance of the FL task. The objective of *PASTEL* is to provide the best trade-off in terms of privacy/utility : (i) Privacy : mitigate membership inference attack by limiting the information shared with the server (ii) Utility : keep the same performance for the local models.
PASTEL is client side mitigation mechanism, *i.e.*, the entire process is fully disclosed to the FL server. *PASTEL* is designed to reduce the generalization gap, which is the difference between the model's performance on the training data and its performance on unseen data. This gap can be exploited by an adversary to infer whether a particular record was used during the training of the model, which can compromise the privacy of the individual. The workflow of *PASTEL* is illustrated in Figure 2. During the training process, each FL client considers minimizing the loss function based on the model output and the real label, and the *Jensen-Shannon Divergence* loss in order to improve the model's ability to generalize to unseen data, while also making it harder for an attacker to infer membership information. Adaptative Gradient Descent (AdaGrad) is used to train the client models locally by ensuring a satisfactory accuracy for the the main classification task. AdaGrad adapts the learning rate for each parameter based on the history of its gradient. This helps to prevent overfitting and distortion of the model by providing a more fine-grained update scheme for each parameter.

2.1. Design Principles of *PASTEL*

Membership inference attacks (MIAs) rely on the overfitting of deep learning model [20]. Intuitively, the generalization gap has been used to mount MIAs, and [11] shows a strong correlation between them. In particular a model with large generalization gap is more vulnerable towards

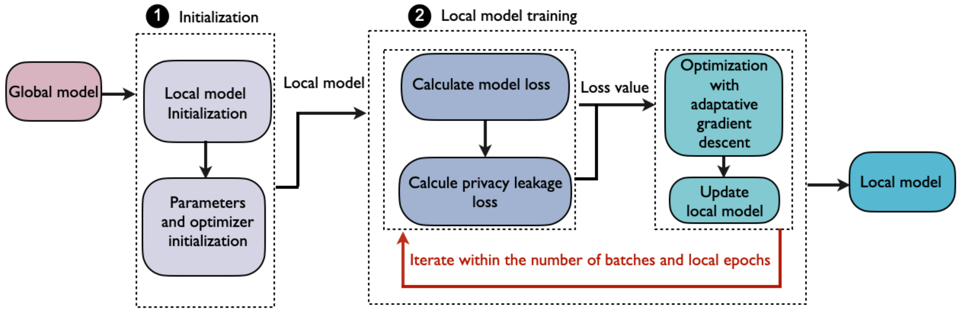


FIGURE 2 – *PASTEL* pipeline at the client-side

MIAs. A model generalization gap g is defined to be $g = a_M - a_{NM}$. where a_M is the model accuracy on training data, *i.e.*, member data and a_{NM} is the model's accuracy on on a dataset drawn from the same distribution as the training data *i.e.*, non-member data. Moreover, [23] shows that the generalization gap on the hidden layers, defined as internal generalization gap, is more important than the output layer. The internal generalization gap is measured based on the divergence of member features and non-member features on the hidden layers. *PASTEL* focuses on the internal generalization gap. To evaluate the distribution shift between member and non-member features, we use the Jensen-Shannon divergence (JSD) [16], a widely used measure of similarity between probability distributions, proven to be robust and less affected by outliers and noise than other distance measures such as Euclidean distance or cosine similarity.

More formally, considering a model M with n layers $\{l_1, l_2, ..., l_n\}$, $\{G_{Xl_1}, G_{Xl_2}, ..., G_{Xl_n}\}$ the gradients of the layer l_i of batch X of member data, $\{G_{\widehat{X}l_1}, G_{\widehat{X}l_2}, ..., G_{\widehat{X}l_n}\}$ the gradients of the layer l_i of batch $\widehat{X}$ of non-member data , the goal of *PASTEL* is to minimize the Jensen-Shannon divergence between the gradients of hidden layers between member and non-member data. JSD is a smoothed and symmetrized version of Kullback-leibler divergence (KLD) [8], and it's calculted based on the latter. to compute the JSD, first, we compute the KLD between member and non-member gradient distributions and then the KLD between non-member and member gradient distributions as defined in Eq. 1 and 2,respectively :

$$KL_m = \mathrm{KL}(G_{\widehat{X}l_i} || \frac{G_{Xl_i} + G_{\widehat{X}l_i}}{2}) \tag{1}$$

$$KL_{nm} = \mathrm{KL}(G_{Xl_i} || \frac{G_{Xl_i} + G_{\widehat{X}l_i}}{2}) \tag{2}$$

The JSD is the average of the two KLD distances defined in the previous equations. 1 and 2 and is computed as follows :

$$\min_{i \in n} \mathrm{JS}(G_{Xl_i} || G_{\widehat{X}l_i}) = \frac{1}{2}(KL_m + KL_{nm}) \tag{3}$$

And the $\mathrm{KL}(\hat{y}||y)$ is measured as follows :

$$\mathrm{KL}(\hat{y}||y) = \sum_{c=1}^{M} \hat{y}_c \log \frac{\hat{y}_c}{y_c} \tag{4}$$

The model loss is calculted based on the JSD loss and labels loss, which is the loss between the model output and the real label Y, as definded in Eq 5.

$$L = L_{jsd} + \mathcal{L}(Y, M(X)) \tag{5}$$

The detailed algorithm of *PASTEL* is definded in algorithm 1. The proposed algorithm follows a multi-step process to optimize the model's performance. Firstly, the algorithm assesses how well the model is currently performing on the training data and identifies misclassified examples. Next, the algorithm calculates the label loss to determine the model's accuracy on the labeled training data and aims to improve it. Finally, the algorithm measures the similarity between the probability distributions of the training and validation data using the Jenson-Shannon Divergence (JSD) loss to improve the model's generalization ability. By optimizing the model's performance based on these insights, the algorithm aims to improve its accuracy and generalization ability. This approach offers a novel way to optimize the performance of deep learning models by focusing on both the accuracy on the training data and the generalization ability to unseen data. The results of our experiments demonstrate the effectiveness of the proposed algorithm in improving the performance of the model on a wide range of datasets.

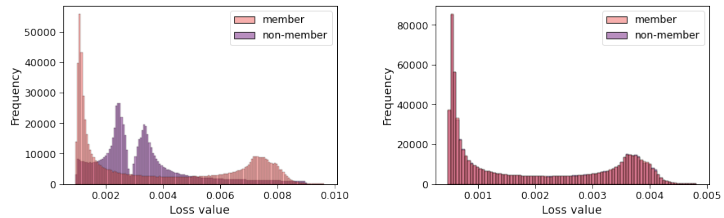
Algorithm 1 *PASTEL* algorithm : $\mathcal{B} \times \mathcal{V} \times \mathcal{W}_t \to \mathcal{W}_{i,t+1}$

Global Model $\mathcal{W}_t$
Loss Function $\mathcal{L}$
Training Batch $(\mathcal{B}, \mathcal{Y}) = \{(B_1, Y_1), \ldots, (B_x, Y_x)\}$
Validation Batch $\mathcal{V} = \{V_i, \ldots, V_x\}$
Local Epochs $\mathcal{E}$
// Initilization
$\mathcal{W}_{i,t+1} = \mathcal{W}_t$
for epoch $\in \mathcal{E}$ **do**
  for (B_i, Y_i) in $(\mathcal{B}, \mathcal{Y})$ **do**
    // Perform forward pass
    $\widehat{Y}_i = W_t(B_i)$
    // Compute model loss
    $l_{label} = \mathcal{L}(Y_i, \widehat{Y}_i)$
    // Compute JSD loss
    $l_{JSD} = JS(B_i, V_i)$
    // Compute gradient
    ∇_t = *AdaGrad*$((l_{label}, l_{JSD}), \mathcal{W}_t)$
    // Update local model
    $\mathcal{W}_{i,t+1} = \mathcal{W}_{i,t+1} + \nabla\, \mathcal{W}_{i,t+1}$
  end for
end for

3. Experimental Evaluation

Generalization Gap Reduction with *PASTEL*. *PASTEL* is a regularization method that aims to reduce the internal generalization gap between member and non-member data. This is achieved by adding a regularization term to the loss function that encourages the model to produce similar outputs for member and non-member inputs. By doing so, *PASTEL* helps to prevent the model from overfitting to the member data and leaking sensitive information.
In our study, we applied *PASTEL* to Texas100 and plotted the resulting loss histograms on member and non-member data in 3. As shown in the figure, *PASTEL* blurs the shift between member and non-member loss distributions, indicating a reduction in the internal generalization gap, which naturally leads to a internal generalization gap and reduced privacy leakage [25], [4].
To further enhance the effectiveness of *PASTEL*, we used adaptive gradient descent to train the

model and minimize the internal generalization gap. This combination of techniques resulted in a more robust and privacy-preserving model that can better handle real-world scenarios.
Overall, our results demonstrate the effectiveness of *PASTEL* in reducing the internal generalization gap and improving the privacy-preserving capabilities of machine learning models.

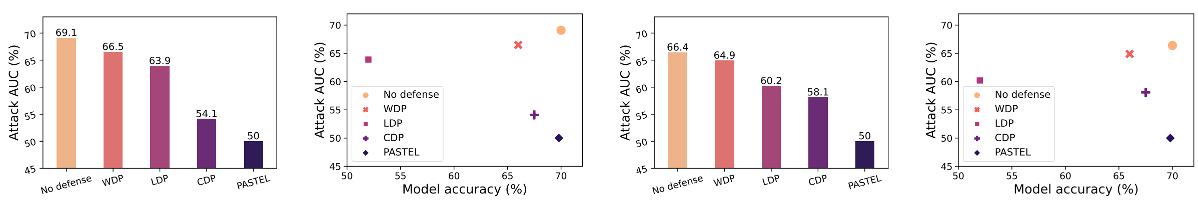


(a) Without *PASTEL* (b) with *PASTEL*

FIGURE 3 – Loss histogram with Purchase100 with fully connected neural network

3.1. Comparison of *PASTEL* Against Other FL Privacy Preserving Mechanisms

We compare *PASTEL* against white box membership inference attack described in [20]. We use Cifar-10 as a benchamrk dataset, and we evaluate *PASTEL*, *WDP*, *LDP*, *CDP* by considering both the utility and membership inference attack AUC on local and global models.
In Figure 4 , we report the results for the datasets Cifar-10, we evaluate the Attack AUC on the global and local model, and the privacy-utility trade-off (the mechanisms with the best trade-off are on the bottom-right corner). State-of-the-art systems *WDP* and *LDP* present moderate mitigation rates (e.g : for Cifar-10, *WDP* reduce the attack AUC of only 2% on the global model and 3% on the local models as showed in the Figures), however, they induce a negative impact on the model utility, for instance, *CDP* reduce the attack AUC of only 6% and the model utility of almost 20%. *PASTEL* allows to mitigate the attack on the different datasets and presents the most competitive results. The AUC attack doesn't go far from 50% on the local model or the global models while maintaining a fairly high model accuracy rate equal to the baseline one.



(a) Attack success on local models (b) Privacy vs. utility of local models (c) Attack success on global model (d) Privacy vs. utility of global model

FIGURE 4 – Privacy leakage with *PASTEL* and state-of-art protection mechanisms – Cifar-10 with ResNet (CNN Model)

4. Conclusion

The paper presents *PASTEL*, a defense mechanism against membership inference attacks in federated learning that minimizes internal generalization gap in neural networks. *PASTEL* provides effective protection against attacks without damaging model utility or breaking secure aggregation guarantees. The authors validate *PASTEL* on various datasets and show that it outperforms other defense mechanisms. The solution achieves its goals of confidentiality and utility without significant extra-computational time or resources. Future work includes exploring *PASTEL*'s performance against other privacy threats, using metrics to parameterize it, and investigating its adaptation in multi-objective frameworks.

References

1. Arivazhagan (M. G.), Aggarwal (V.), Singh (A. K.) et Choudhary (S.). – Federated Learning with Personalization Layers. *arXiv :1912.00818*, vol. 1912.00818, 2019.
2. Bonawitz (K. A.), Ivanov (V.), Kreuter (B.), Marcedone (A.), McMahan (H. B.), Patel (S.), Ramage (D.), Segal (A.) et Seth (K.). – Practical Secure Aggregation for Privacy-Preserving Machine Learning. – In Thuraisingham (B. M.), Evans (D.), Malkin (T.) et Xu (D.) (édité par), *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, October 30 - November 03, 2017*, pp. 1175–1191, Dallas, TX, USA, 2017. ACM.
3. Cavallaro (L.), Kinder (J.), Afroz (S.), Biggio (B.), Carlini (N.), Elovici (Y.) et Shabtai (A.) (édité par). – *HybridAlpha : An Efficient Approach for Privacy-Preserving Federated Learning*. – ACM, 2019.
4. Chen (D.), Yu (N.) et Fritz (M.). – Relaxloss : Defending membership inference attacks without losing utility. – In *The Tenth International Conference on Learning Representations, ICLR 2022, April 25-29, 2022*, Virtual Event, 2022. OpenReview.net.
5. Haddadpour (F.), Kamani (M. M.), Mokhtari (A.) et Mahdavi (M.). – Federated learning with compression : Unified analysis and sharp guarantees. – In Banerjee (A.) et Fukumizu (K.) (édité par), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021*, *Proceedings of Machine Learning Research*, volume 130, pp. 2350–2358, Virtual Event, 2021. PMLR.
6. Hidano (S.), Murakami (T.), Katsumata (S.), Kiyomoto (S.) et Hanaoka (G.). – Model Inversion Attacks for Online Prediction Systems : Without Knowledge of Non-Sensitive Attributes. *IEICE Trans. Inf. Syst.*, vol. 101-D, n11, 2018, pp. 2665–2676.
7. Hitaj (B.), Ateniese (G.) et Pérez-Cruz (F.). – Deep models under the GAN : information leakage from collaborative deep learning. – In Thuraisingham (B.), Evans (D.), Malkin (T.) et Xu (D.) (édité par), *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, CCS 2017, October 30 - November 03, 2017*, pp. 603–618, Dallas, TX, USA, 2017. ACM.
8. Joyce (J. M.). – *Kullback-Leibler Divergence*, pp. 720–722. – Berlin, Heidelberg, Springer Berlin Heidelberg, 2011.
9. Kairouz (P.), McMahan (H. B.), Avent (B.), Bellet (A.), Bennis (M.), Bhagoji (A. N.), Bonawitz (K. A.), Charles (Z.), Cormode (G.), Cummings (R.), D'Oliveira (R. G. L.), Rouayheb (S. E.), Evans (D.), Gardner (J.), Garrett (Z.), Gascón (A.), Ghazi (B.), Gibbons (P. B.), Gruteser (M.), Harchaoui (Z.), He (C.), He (L.), Huo (Z.), Hutchinson (B.), Hsu (J.), Jaggi (M.), Javidi (T.), Joshi (G.), Khodak (M.), Konečný (J.), Korolova (A.), Koushanfar (F.), Koyejo (S.), Lepoint (T.), Liu (Y.), Mittal (P.), Mohri (M.), Nock (R.), Özgür (A.), Pagh (R.), Raykova (M.), Qi (H.), Ramage (D.), Raskar (R.), Song (D.), Song (W.), Stich (S. U.), Sun (Z.), Suresh (A. T.), Tramèr (F.), Vepakomma (P.), Wang (J.), Xiong (L.), Xu (Z.), Yang (Q.), Yu (F. X.), Yu (H.) et Zhao (S.). – Advances and Open Problems in Federated Learning. *ArXiv*, vol. abs/1912.04977, 2019.
10. Konečný (J.), McMahan (H. B.), Ramage (D.) et Richtárik (P.). – Federated Optimization : Distributed Machine Learning for On-Device Intelligence. *ArXiv*, vol. abs/1610.02527, 2016.
11. Li (J.), Li (N.) et Ribeiro (B.). – Membership inference attacks and defenses in supervised learning via generalization gap. *ArXiv*, vol. abs/2002.12062, 2020.
12. Liu (P.), Xu (X.) et Wang (W.). – Threats, attacks and defenses to federated learning : issues, taxonomy and perspectives. *Cybersecur.*, vol. 5, n1, 2022, p. 4.
13. Malekzadeh (M.), Clegg (R. G.), Cavallaro (A.) et Haddadi (H.). – Mobile sensor data anonymization. – In *Proceedings of the International Conference on Internet of Things Design and Implementation*, *IoTDI '19*, IoTDI '19, pp. 49–58, New York, NY, USA, 2019. ACM.

14. McMahan (B.), Moore (E.), Ramage (D.), Hampson (S.) et y Arcas (B. A.). – Communication-Efficient Learning of Deep Networks from Decentralized Data. – In Singh (A.) et Zhu (X. J.) (édité par), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017,*, *Proceedings of Machine Learning Research*, volume 54, pp. 1273–1282, Fort Lauderdale, FL, USA, 2017. PMLR.
15. McSherry (F.). – Privacy integrated queries : an extensible platform for privacy-preserving data analysis. *Commun. ACM*, vol. 53, n9, 2010, pp. 89–97.
16. Menéndez (M.), Pardo (J.), Pardo (L.) et Pardo (M.). – The jensen-shannon divergence. *Journal of the Franklin Institute*, vol. 334, n2, 1997, pp. 307–318.
17. Naseri (M.), Hayes (J.) et Cristofaro (E. D.). – Toward robustness and privacy in federated learning : Experimenting with local and central differential privacy. *ArXiv*, vol. abs/2009.03561, 2020.
18. Nasr (M.), Shokri (R.) et Houmansadr (A.). – Comprehensive Privacy Analysis of Deep Learning : Passive and Active White-box Inference Attacks against Centralized and Federated Learning. – In *IEEE Symposium on Security and Privacy, SP 2019*, pp. 739–753, San Francisco, CA, USA, mai 2019. IEEE.
19. Salem (A.), Zhang (Y.), Humbert (M.), Berrang (P.), Fritz (M.) et Backes (M.). – ML-Leaks : Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. – In *26th Annual Network and Distributed System Security Symposium, NDSS 2019, February 24-27, 2019*, San Diego, California, USA, 2019. The Internet Society.
20. Shokri (R.), Stronati (M.), Song (C.) et Shmatikov (V.). – Membership Inference Attacks Against Machine Learning Models. – In *IEEE Symposium on Security and Privacy, S&P 2017*, pp. 3–18, San Jose, CA, USA, May 2017. IEEE.
21. Sun (L.), Qian (J.) et Chen (X.). – LDP-FL : Practical Private Aggregation in Federated Learning with Local Differential Privacy. – In Zhou (Z.) (édité par), *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, 19-27 August 2021*, pp. 1571–1578, Virtual Event / Montreal, Canada, 2021. ijcai.org.
22. Wang (H.), Yurochkin (M.), Sun (Y.), Papailiopoulos (D. S.) et Khazaeni (Y.). – Federated Learning with Matched Averaging. – In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, Addis Ababa, Ethiopia, Apr 2020. OpenReview.net.
23. Wu (D.), Qi (S.), Qi (Y.), Li (Q.), Cai (B.), Guo (Q.) et Cheng (J.). – Understanding and defending against white-box membership inference attack in deep learning. *Knowl. Based Syst.*, vol. 259, 2023, p. 110014.
24. Yang (Q.), Liu (Y.), Chen (T.) et Tong (Y.). – Federated Machine Learning : Concept and Applications. *ACM Transactions on Intelligent Systems and Technology*, vol. 10, n2, 2019.
25. Yeom (S.), Giacomelli (I.), Fredrikson (M.) et Jha (S.). – Privacy risk in machine learning : Analyzing the connection to overfitting. – In *31st IEEE Computer Security Foundations Symposium, CSF 2018, July 9-12, 2018*, pp. 268–282, Oxford, United Kingdom, 2018. IEEE Computer Society.
26. Zhu (L.) et Han (S.). – Deep leakage from gradients. *Springer*, vol. 12500, 2020, pp. 17–31.