Inference Attacks on Federated Learning - A Survey

Zohar Cochavi

Abstract—Federated Learning (FL) is a privacy-preserving approach to distributed machine learning, but it is vulnerable to inference attacks. Inference attacks aim to extract information about regarding model or data used in FL, directly threatening one of its core principles. This essay provides an overview of state-of-the-art inference attacks in FL and their implications for data privacy. It introduces the basics of FL, different types of inference attacks (model inversion, membership inference, property inference), and provides an overview of recent research on gradient inversion and membership inference attacks. We emphasize the need for robust defenses to safeguard sensitive information in FL systems, along with the importance of future research in addressing these increasingly realistic threats.

Introduction

Machine learning models have demonstrated remarkable capabilities in interpreting and deriving insights from data, leading to significant advancements in fields such as medical diagnostics (Kononenko 2001) and natural language processing (Liu et al. 2023). However, these applications often involve handling privacy-sensitive data (Rieke et al. 2020), which can be inferred from trained models (Fredrikson, Jha, and Ristenpart 2015), raising concerns about data privacy and security. Federated Learning is a privacy-preserving approach to distributed machine learning that aims to address these concerns (McMahan and Ramage 2017) but is not immune to potential exploits (Abad et al. 2022). Inference Attacks directly threaten this privacy-preserving feature by attempting to extract information about the model and the data it was trained on (Geiping et al. 2020).

In this essay, I provide an overview of state-of-the-art Inversion Attacks and their defenses to Federated Learning. Besides informing on the state-of-the-art, this essay should provide an introduction to the subject for both machine learning and cyber security specialists. First, we will discuss the basics of Federated Learning, introduce attacks on Machine Learning models, and consider a taxonomy of Inference Attacks on Federated Learning. Then, I will present novel attacks and some of their defenses. Finally, we will discuss the threat these attacks present when considering the defenses and present future work to accommodate for these threats.

Background

In this section, the necessary background information will be introduced. The background is considered whatever already existed until last year (March 2022). We will provide a concise overview of machine learning principles, to then discuss the workings of Federated Learning (FL). Having covered the necessary machine learning knowledge, the discussion will move to how one would attack such systems. Finally, we focus on previous inference attacks as summarized and discussed by (Abad et al. 2022).

Machine Learning

The goal of any machine learning algorithm is to predict some label or value given familiar but unseen data. For the purposes of this discussion, the machine-learning process can be separated into 3 stages:

1. Training
2. Testing/Evaluation
3. Deployment

During training, the machine learning model, f is given a set of tuples $\{(x_i, y_i)\}$. The learning algorithm then adjusts the model parameters, θ, such that the model after training, f_θ, maps the input features x to the target value(s) y. Depending on the learning task, y could be a continuous value (regression), a binary value (binary classification), or a set of discrete values (multi-class classification)[1] (Abad et al. 2022; Chakraborty et al. 2018).

The testing phase assesses the performance of the model. We take a similar, but unseen set of tuples $\{(x, y)\}$ and test whether the model, $f_\theta(x)$, returns the correct value(s). It's important to only test on *unseen* data since the aim is to assess the *generalizing* ability of the model. One can imagine the performance to be higher if the same data that was used to train, was used to test.

After the model is trained and evaluated, it is then deployed. Often accessed via a public or private API. In the case of Federated Learning, this process is iterative.

Federated Learning

Federated Learning (FL) is a method of delegating, or democratizing, the training stage of a machine learning algorithm. Its benefits are threefold (Konečný et al. 2017):

1. It avoids sharing *raw* personal data with a third party.
2. Processing resources are delegated.
3. Data that is fragmented over multiple locations can still be used for training.

1. More types of machine learning exist (Oprea and Vassilev 2023), but the details of their taxonomy are not important for this discussion. The focus will be on Federated Learning which, almost exclusively uses supervised learning.

The process, generally, works as follows. Each client in set of clients, $C = \{c_0, \ldots, c_n\}$, trains a private machine learning model their respective dataset. Information about this trained model is then sent to a central server that *aggregates* the information from all clients into a single model. The newly trained model is then sent back to the clients for another iteration[2] Konečný et al. (2017).

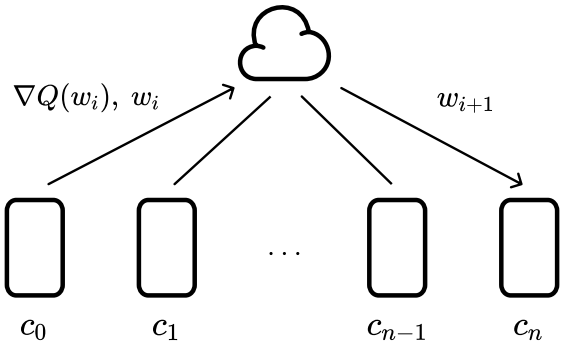

Figure 1: Typical Federated Learning network topology. The client, c_i, sends the gradient, $\nabla Q(\theta_i)$, and/or weights, θ_i, of a particular iteration i. The central server then sends the updated model parameters θ_{i+1} initiating the next iteration.

Various aggregation algorithms exist. The most popular of which are *Federated Stochastic Gradient Descent* (FedSGD) and *Federated Averaging* (FedAvg). These use the *gradient* (the function that describes how to optimize the model) or the aforementioned parameters, θ, of the client models respectively when communicating model updates Gu, Bai, and Xu (2022). While technical details of these algorithms are not crucial to this discussion, it is important to note that both the gradient and the weights contain embedded information about the client's dataset (Fredrikson, Jha, and Ristenpart 2015; Geiping et al. 2020).

Vertical versus Horizontal FL

Lastly, there are two types of FL: Horizontal Federated Learning (HFL), or cross-device Federated Learning; and Vertical Federated Learning (VFL), or cross-silo Federated Learning (Suri et al. 2022; Abad et al. 2022). These differentiate in the way the client data is aligned to provide a trainable model.

In the former, devices all collect data on the same features, but their sample space is not equal (the distributions might not align, and the size can be different). This is the most used type of Federated Learning, which also is how companies such as Google train their models on user data (McMahan and Ramage 2017).

In the latter, we can imagine a set of hospitals or companies (or *data silos*) that need to train a model on *all* the available user data. The data, however, cannot be directly shared and the kind of data they collect on their users is also different. Regardless, each user, or *subject*, is present in each database. They thus share the same sample space but differ in the features they train their local models on.

2. Other network topologies are possible within Federated Learning. One notable example is a peer-to-peer network which allows for a completely decentralized machine-learning approach. Most actual implemented systems, however, have a normal client-server structure (Abad et al. 2022).

Attacking Machine Learning

The machine learning approaches discussed so far (normal/*centralized* learning and Federated Learning) contain several points at which an attacker could intervene to exploit various characteristics of the system. Before discussing inference attacks that take place in later stages of the machine learning pipeline, let us briefly discuss other potential threats to machine learning models.

The phases discussed in the first section (training, testing, and deployment) correspond directly to different attacks which can be categorized as follows (Chakraborty et al. 2018).

1. *Poisoning Attack*: This type of attack, known as contamination of the training data, takes place during the training time of the machine learning model. An adversary tries to poison the training data by injecting carefully designed samples to compromise the whole learning process eventually.
2. *Evasion Attack*: This is the most common type of attack in the adversarial setting. The adversary tries to evade the system by adjusting malicious samples during the testing phase. This setting does not assume any influence over the training data.
3. *Inference/Exploratory Attack*: These attacks do not influence any dataset. Given black-box access to the model, they try to gain as much knowledge as possible about the learning algorithm of the underlying system and pattern in training data.

While the first two also pose potential threats to the FL scheme and are very popular in centralized machine learning, they are considerably harder to perform on Federated Learning as the data is distributed (Tolpegin et al. 2020). Databases of multiple clients have to be compromised to create an exploit that is comparable to that of a centralized machine-learning approach.

Inference attacks, however, threaten the privacy guarantees FL attempts to give. Inference attacks specifically try to *infer* information about the dataset the model was trained on or the model itself. Thereby threatening the confidentiality of the database, and thus the privacy, of the victims involved (Abad et al. 2022). Since they actively infer information about a deployed system, the amount of information on the system determines how powerful such an attack could be. For this reason, they are also further specified as *white-box* or *black-box*, and sometimes *grey-box* inference attacks (Nasr, Shokri, and Houmansadr 2019).

Inference Attacks

Inference attacks can be applied to both centralized machine learning models and Federated Learning schemes. Many of the principles we will cover apply to both centralized and Federated Learning, but the focus will be on applications on FL. Specifically, we will provide an overview of attack classifications as given by Abad et al. (2022).

Firstly, depending on the target information the attacker attempts to infer, the attack is classified as follows:

- *Model Inversion*: In model inversion, the attacker attempts to invert the machine learning model. Thereby finding the data point corresponding to a certain label. Fredrikson, Jha, and Ristenpart (2015) were able to invert a facial recognition model, allowing them to recover the image of an individual known to be in the training dataset.
- *Membership Inference*: In this attack, the goal is to determine whether a data point (x, y) was part of the training set. In FL it is also possible to determine whether a data point was part of the training set of a particular client.
- *Property Inference*: Property inference attempts to detect whether the dataset used to train the (local) model contains some property (e.g. whether images are noisy) (Ganju et al. 2018).

Secondly, when considering a malicious central server, the attack can be classified according to the manner in which the attacker interferes with the training procedure (Abad et al. 2022):

- *Passive*: Also known as an *honest-but-curious* scenario. The attacker can only read or eavesdrop on communications (i.e. the weights and gradient transmitted between the clients and the server), and the local model and dataset in case of an honest-but-curious client.
- *Active*: The attack is more effective than the former but less stealthy. It essentially changes the learning goal from minimizing loss to maximizing inferred information from the victim.

Lastly, attacks can be categorized based on the position of the attacker in the network:

- *Local*: The attacker is a client, i.e. they can only access *their* database, parameters, and the global parameters they receive from the server.
- *Global*: The attacker is the central server. They do not have access to any databases but can access the gradients/parameters sent by all clients and the global model.

Inference Attacks in Federated Learning

This section will discuss the state-of-the-art of inference attacks on Federated Learning. Specifically, we will discuss progress in the field as of March 2022. The research presented here was found primarily by querying Google Scholar with the terms "Inference Attacks on Federated Learning", "Membership Inference on Federated Learning", "Model Inversion on Federated Learning", and "Gradient Inversion on Federated Learning". The next section will cover the threats these advances pose to current systems.

First, we will discuss various attacks, focusing not only on their performance but paying special attention to the scenario in which the researchers placed the hypothetical adversary. Then, we will cover defenses to some of these attacks.

Attacking

Various types of attacks fall under the umbrella of inference attacks. As of the writing of this essay, the most popular are *Membership Inference* and *Gradient Inversion* as these show the most results. *Passive* inference attacks are more often covered than their *active* counterparts. For each paper, we annotate the type of attack (see Inference Attacks), summarize the findings of the authors, and briefly discuss them.

Do Gradient Inversion Attacks Make Federated Learning Unsafe?

Keywords: *Model Inversion*, *Local/Global*, *Cross-Device/HFL*, *Passive*

Hatamizadeh et al. (2023) performed image reconstruction using gradient inversion while relaxing a strong assumption made in prior work regarding Batch Normalization (BN) (Ioffe and Szegedy 2015). BN is a technique used in neural networks that significantly improve the learning rate and stability, and is therefore ubiquitous in modern machine learning. The technique introduces two learned parameters, β and γ, which thus change during the learning process(Ioffe and Szegedy 2015). Previous work has assumed these statistics to be static (Geiping et al. 2020; Kaissis et al. 2021), introducing an error that would compound over time. The authors were able to reliably reconstruct images without assuming static BN statistics. The authors make a strong case for an inversion attack that could be used in practice but still rely on priors (approximations of the image) to make accurate reconstructions.

Improved Gradient Inversion Attacks and Defenses in Federated Learning

Keywords: *Membership Inference*, *Local/Global*, *Cross-Device/HFL*, *Passive*, *White-Box*

Geng et al. (2023) proposed a framework for inverting both *FedAVG*-based and *FedSGD*-based networks in an "honest-but-curious" scenario. They mention prior work has failed to effectively perform gradient inversion when FL uses the *FedAVG* aggregation algorithm. Furthermore, they specify methods for fine-tuning the performance of image restoration in the inverted model, allowing them to restore images that were introduced 16 epochs before the current iteration. As Federated Learning is an iterative process, one can imagine that the further a data point is removed from the current iteration, the harder it is to infer from the current gradient. While their results are promising, they do assume a white-box attack scenario making their attack harder to perform.

CS-MIA: Membership Inference Attack Based on Prediction Confidence Series in Federated Learning

Keywords: *Membership Inference*, *Local/Global*, *Cross-Device/VFL*, *Passive*

Gu, Bai, and Xu (2022) were able to determine whether data points are members of certain datasets by following the trend in their classification confidence. Over time, the global model should perform less well on participants' private data, meaning that member data should follow a different trend compared to non-member data. They then train a supervised model the determine whether data points were part of the training set based on this assumption. The model is then used to determine

the probability of unseen data being part of the target training data set. They show high accuracy and F1-scores for all datasets with the lowest performer being MNIST (around 60% compared to >90% for the other datasets). Still, the proposed solution scores the best out of all included approaches by a significant margin.

Subject Membership Inference Attacks in Federated Learning

Keywords: *Membership Inference*, *Local/Global*, *Cross-Silo/VFL*, *Passive*

In a black-box setting, Suri et al. (2022) propose a method for what they call *Subject Inference* (see Federated Learning). They describe previous work as being disconnected from real-world scenarios as it (i) includes information adversaries would not normally have access to and (ii) assumes the adversary is looking for data points rather than individuals. Instead of determining whether one particular data point was part of the training set, the authors attempt to infer whether an individual, a *subject*, (or rather their distribution) is present in the dataset given some preexisting information on them. They show the attack to be very effective in various real-world datasets while also increasing the realism of the scenario. They show Subject Inference to be a real threat to user privacy.

Active Membership Inference Attack under Local Differential Privacy in Federated Learning.

Keywords: *Membership Inference*, *Local/Global*, *Cross-Device/HFL*, *Active*

Different from other works, Nguyen et al. (2023) considers a maximally malicious, i.e. *active*, membership inference attack. They implement a method for inferring membership of a particular data point in the presence of differential privacy (Dwork and Roth 2013). Differential privacy obscures the relation of the individual to the data point, without affecting the patterns used for training machine learning models. The authors show that their method performs well, even under such obscuring of the data. Furthermore, the attack only starts to degrade after the level of obscurity interferes with model performance. They show that more rigorous privacy methods should be proposed to deal with such attacks.

Defending

To combat inference attacks, we discuss potential defenses against them. Some of the papers that are included have been discussed in the last section. These have been marked with a footnote accordingly [3]. For each paper, we will summarize the proposed measures and briefly discuss them.

Improved Gradient Inversion Attacks and Defenses in Federated Learning[4]

Geng et al. (2023) found that labels that only appeared only once were more prone to their proposed inversion attacks (see Attacks). They also mention the use of larger batch sizes in the global model (i.e. more clients) to reduce the amount of private information embedded in a single batch. Lastly, they claim FedAVG possesses "stronger privacy preserving capabilities than FedSGD". As this was included in the discussion of their attack-oriented paper, they do not evaluate these claims further.

Do Gradient Inversion Attacks Make Federated Learning Unsafe?[5]

Hatamizadeh et al. (2023) make several recommendations to make existing implementations of FL safer, namely: (i) larger training sets, (ii) updates from a larger number of iterations over different (iii) large batch sizes. In addition, they mention three more changes that could potentially mitigate server-side (i.e. *Global*) gradient inversion attacks: (1) The use of *Homomorphic Encryption* (see Future Work), (2) ensuring the attacker does not have knowledge of the model architecture, and (3) using an alternative aggregation algorithm such as FedBN Andreux et al. (2020). The countermeasures provided are relatively general. They also provided sources affirming their suspicions.

An Empirical Analysis of Image Augmentation Against Model Inversion Attack in Federated Learning

Shin et al. (2023) propose the use of image augmentation as a more viable alternative to differential privacy (Dwork and Roth 2013). Image augmentation is a data synthesis method that increases the size of the training set, and reduces over-fitting (Shorten and Khoshgoftaar 2019). As this introduces fake data while improving the overall performance of the model, the authors suggest it could be used to mitigate model inversion attacks. The attack they used was introduced by (Geiping et al. 2020), and various more successful attacks have been constructed since then Geng et al. (2023).

ResSFL: A Resistance Transfer Framework for Defending Model Inversion Attack in Split Federated Learning

In a framework introduced by (J. Li et al. 2022), Split Federated Learning (SFL) (Annavaram and Avestimehr, n.d.) is augmented with a model that attempts to invert the model before the client sends their model to the central server. By choosing weights where the discriminator performs poorly, they claim to improve the resiliency of the scheme to model inversion attacks. They indeed show improvements over the standard implementation of SFL but do not mention how this method compares to attacks on default FL.

Discussion

In this section, we will discuss the attacks and defenses as presented in the last section. Specifically, we determine the threats these attacks pose and if the defenses included could effectively mitigate the. In the end, we propose new research directions that could help mitigate these threats.

Current Threats and Trends

The attacks presented show how Federated Learning might not be able to guarantee privacy. Privacy, thus, should still remain a concern even if Federated Learning brings stronger

3. Often, novel attack proposals also include possible countermeasures. Some of the papers covered in the last section, therefore, have also been included in this section.

4. Often, novel attack proposals also include possible countermeasures. Some of the papers covered in the last section, therefore, have also been included in this section.

5. Often, novel attack proposals also include possible countermeasures. Some of the papers covered in the last section, therefore, have also been included in this section.

privacy guarantees than traditional machine learning and its derivatives. Let us summarize the threats these attacks pose:

- *More Realistic Scenarios*: Research starts to introduce more realistic scenarios that could threaten current implementations of Federated Learning. As the field matures, attacks seem to become more realistic. Especially the work presented by Suri et al. (2022) poses a real threat as it assumes a complete black-box attack with reasonable assumptions while still showing good performance. Even in complete black-box settings, however, we still assume the ability to intercept and read the communications. Were this to be encrypted, such attacks could possibly be mitigated (Y. Li et al. 2021).
- *Increased Resilience Against Existing Privacy Measures*: Some of the aforementioned papers has shown improvements concerning the evasion of privacy-preserving measures. Nguyen et al. (2023) have shown how a membership inference attack can be effectively performed in the presence of differential privacy. Their method was effective to such a degree that the attack was ineffective only once the privacy measures started to affect model performance. The image augmentation countermeasure proposed by (Shin et al. 2023) could be a viable option. This countermeasure, however, was only tested in a *passive* scenario.
- *Stronger Attacks in Existing Scenarios*: As to be expected, some work was focused on improving performance in existing scenarios. Gu, Bai, and Xu (2022) have shown that much is still to be learned in the field by proposing a relatively simple approach that improves upon all previous methods by a large margin.

Such developments are not surprising, progress in both offense and defense is to be expected. The speed at which research moves forward is very impressive and suggests the field is still in the early stages of development. When considering using such new technologies in production, this could be considered when assessing the security of such systems.

Future Work

Considering the aforementioned advances, the following directions could provide useful for future research:

1. Consider using existing preprocessing methods for privacy preservation. Shin et al. (2023) and Hatamizadeh et al. (2023) both either use or suggest using existing pre-processing or other learning-enhancing augmentations to improve privacy. Efforts toward generalizing data *before* training might prove a solution to both overfitting and privacy.
2. New attack methods would benefit from relaxing assumptions instead of attempting to increase performance. By doing this, various of the attacks shown have been to provide a more realistic view of the privacy-preserving features of FL. While performance improvements might provide interesting results and insights, focusing efforts on exposing potential *realistic* threats would have a more direct effect on our ability to assess FL from a privacy perspective.
3. Working, safe Homomorphic Encryption (HE) would hamper most of the aforementioned attacks. Being able to encrypt data *before* training a model would make inference attacks completely benign (Lee et al. 2022). Work from the past year, however, was able to infer privacy-sensitive information about the training set regardless of the presence of HE (Y. Li et al. 2021). More research on the robust use of HE could prove a catch-all solution for many of the presented machine-learning attacks.

Conclusion

This essay has provided an overview for security specialists and machine learning specialists to assess the current state of Inference Attacks in Federated Learning. Progress over the last year has shown the field to be advancing quickly. Introducing successful attacks on new, more realistic scenarios, showing the ability to circumvent mature privacy-preserving, measures and improving the performance of existing methods. The developments seen here are concerning given the prevalence of Federated Learning. More research is needed to assess how privacy-preserving Federated Learning actually is and whether additional countermeasures provide enough security to circumvent the apparent threats presented here.

References

Abad, Gorka, Stjepan Picek, Víctor Julio Ramírez-Durán, and Aitor Urbieta. 2022. "On the Security & Privacy in Federated Learning." arXiv. https://arxiv.org/abs/2112.05423.

Andreux, Mathieu, Jean Ogier Du Terrail, Constance Beguier, and Eric W. Tramel. 2020. "Siloed Federated Learning for Multi-centric Histopathology Datasets." In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, edited by Shadi Albarqouni, Spyridon Bakas, Konstantinos Kamnitsas, M. Jorge Cardoso, Bennett Landman, Wenqi Li, Fausto Milletari, et al., 12444:129–39. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-60548-3_13.

Annavaram, Chaoyang He Murali, and Salman Avestimehr. n.d. "Group Knowledge Transfer: Federated Learning of Large CNNs at the Edge."

Chakraborty, Anirban, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. "Adversarial Attacks and Defences: A Survey." arXiv. https://arxiv.org/abs/1810.00069.

Dwork, Cynthia, and Aaron Roth. 2013. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends® in Theoretical Computer Science* 9 (3-4): 211–407. https://doi.org/10.1561/0400000042.

Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. 2015. "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures." In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 1322–33. CCS '15. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/2810103.2813677.

Ganju, Karan, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. 2018. "Property Inference Attacks on Fully Connected Neural Networks Using Permutation Invariant

Representations." In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 619–33. Toronto Canada: ACM. https://doi.org/10.1145/3243734.3243834.

Geiping, Jonas, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. 2020. "Inverting Gradients - How Easy Is It to Break Privacy in Federated Learning?" In *Advances in Neural Information Processing Systems*, 33:16937–47. Curran Associates, Inc.

Geng, Jiahui, Yongli Mou, Qing Li, Feifei Li, Oya Beyan, Stefan Decker, and Chunming Rong. 2023. "Improved Gradient Inversion Attacks and Defenses in Federated Learning." *IEEE Transactions on Big Data*, 1–13. https://doi.org/10.1109/TBDATA.2023.3239116.

Gu, Yuhao, Yuebin Bai, and Shubin Xu. 2022. "CS-MIA: Membership Inference Attack Based on Prediction Confidence Series in Federated Learning." *Journal of Information Security and Applications* 67 (June): 103201. https://doi.org/10.1016/j.jisa.2022.103201.

Hatamizadeh, Ali, Hongxu Yin, Pavlo Molchanov, Andriy Myronenko, Wenqi Li, Prerna Dogra, Andrew Feng, et al. 2023. "Do Gradient Inversion Attacks Make Federated Learning Unsafe?" *IEEE Transactions on Medical Imaging*, 1–1. https://doi.org/10.1109/TMI.2023.3239391.

Ioffe, Sergey, and Christian Szegedy. 2015. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." arXiv. https://doi.org/10.48550/arXiv.1502.03167.

Kaissis, Georgios, Alexander Ziller, Jonathan Passerat-Palmbach, Théo Ryffel, Dmitrii Usynin, Andrew Trask, Ionésio Lima, et al. 2021. "End-to-End Privacy Preserving Deep Learning on Multi-Institutional Medical Imaging." *Nature Machine Intelligence* 3 (6): 473–84. https://doi.org/10.1038/s42256-021-00337-8.

Konečný, Jakub, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2017. "Federated Learning: Strategies for Improving Communication Efficiency." arXiv. https://doi.org/10.48550/arXiv.1610.05492.

Kononenko, Igor. 2001. "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective." *Artificial Intelligence in Medicine* 23 (1): 89–109. https://doi.org/10.1016/S0933-3657(01)00077-X.

Lee, Joon-Woo, Hyungchul Kang, Yongwoo Lee, Woosuk Choi, Jieun Eom, Maxim Deryabin, Eunsang Lee, et al. 2022. "Privacy-Preserving Machine Learning With Fully Homomorphic Encryption for Deep Neural Network." *IEEE Access* 10: 30039–54. https://doi.org/10.1109/ACCESS.2022.3159694.

Li, Jingtao, Adnan Siraj Rakin, Xing Chen, Zhezhi He, Deliang Fan, and Chaitali Chakrabarti. 2022. "ResSFL: A Resistance Transfer Framework for Defending Model Inversion Attack in Split Federated Learning." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10194–202.

Li, Xiaoxiao, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. 2021. "FedBN: Federated Learning on Non-IID Features via Local Batch Normalization." arXiv. https://doi.org/10.48550/arXiv.2102.07623.

Li, Yuchen, Yifan Bao, Liyao Xiang, Junhan Liu, Cen Chen, Li Wang, and Xinbing Wang. 2021. "Privacy Threats Analysis to Secure Federated Learning." arXiv. https://arxiv.org/abs/2106.13076.

Liu, Yiheng, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, et al. 2023. "Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models." *arXiv.org*. https://arxiv.org/abs/2304.01852v3.

McMahan, Brendan, and Daniel Ramage. 2017. "Federated Learning: Collaborative Machine Learning Without Centralized Training Data." https://ai.googleblog.com/2017/04/federated-learning-collaborative.html.

Nasr, Milad, Reza Shokri, and Amir Houmansadr. 2019. "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks Against Centralized and Federated Learning." In *2019 IEEE Symposium on Security and Privacy (SP)*, 739–53. https://doi.org/10.1109/SP.2019.00065.

Nguyen, Truc, Phung Lai, Khang Tran, NhatHai Phan, and My T. Thai. 2023. "Active Membership Inference Attack Under Local Differential Privacy in Federated Learning." arXiv. https://doi.org/10.48550/arXiv.2302.12685.

Oprea, Alina, and Apostol Vassilev. 2023. "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations (Draft)." NIST AI 100-2e2023 ipd. National Institute of Standards and Technology.

Rieke, Nicola, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, et al. 2020. "The Future of Digital Health with Federated Learning." *NPJ Digital Medicine* 3 (September): 119. https://doi.org/10.1038/s41746-020-00323-1.

Shin, Seunghyeon, Mallika Boyapati, Kun Suo, Kyungtae Kang, and Junggab Son. 2023. "An Empirical Analysis of Image Augmentation Against Model Inversion Attack in Federated Learning." *Cluster Computing* 26 (1): 349–66. https://doi.org/10.1007/s10586-022-03596-1.

Shorten, Connor, and Taghi M. Khoshgoftaar. 2019. "A Survey on Image Data Augmentation for Deep Learning." *Journal of Big Data* 6 (1): 60. https://doi.org/10.1186/s40537-019-0197-0.

Suri, Anshuman, Pallika Kanani, Virendra J. Marathe, and Daniel W. Peterson. 2022. "Subject Membership Inference Attacks in Federated Learning." *arXiv.org*. https://arxiv.org/abs/2206.03317v3.

Tolpegin, Vale, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. 2020. "Data Poisoning Attacks Against Federated Learning Systems." *arXiv.org*. https://arxiv.org/abs/2007.08432v2.