Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

Privacy Threats Unveiled: A Comprehensive Analysis of Membership Inference Attacks on Machine Learning Models and Defense Strategies

Ali Sezer ÇAM[1], Fatih YILDIZ[1]*

[1]*Erzurum Technical University, Erzurum, Türkiye, ali.cam97@erzurum.edu.tr*

ABSTRACT

Membership inference attacks, aiming to determine whether target data belongs to a training dataset through machine learning model exploitation, present an escalating privacy threat within the machine learning landscape. This study initiates from fundamental theories surrounding the attack and defense mechanisms of machine learning models. The paper conducts a thorough analysis of key technical models, elucidating the intricate relationship between attack models and potential privacy risks to ensure data privacy security and advance the realm of machine learning applications. The introduction covers the adversary model of membership inference attacks, encompassing definitions, classifications, and the generation mechanism. Additionally, the paper provides a comprehensive overview and analysis of existing membership inference attack algorithms. Practical applications of membership inference attacks are explored, followed by the categorization and comparison of defense techniques. The study concludes with a comparative analysis of existing attack schemes and their corresponding defense technologies, offering insights into the evolving landscape of membership inference attacks in machine learning. The work not only anticipates future research challenges in data privacy protection but also establishes a theoretical foundation crucial for addressing data privacy leakage, thereby significantly contributing to the progress of machine learning applications.

Keywords: Membership Inference Attacks, Security, Machine Learning, Defense Strategies, Data Privacy

1. INTRODUCTION

The rapid evolution of artificial intelligence, particularly machine learning theory and technology, owes much to the internet's progress, hardware updates, extensive data collection, and the advancement of intelligent algorithms [1]. Its widespread application in diverse fields, including data mining [2], computer vision [3] [4], email filtering [5], credit card fraud detection [6] [7] [8] [9], and medical diagnosis [10] [11], has significantly enhanced efficiency through the analysis of large datasets. Despite the convenience and intelligence offered by machine learning, the increased collection of personal sensitive information, such as physiological characteristics, medical records, and social networks, has introduced severe challenges to the security and privacy of this burgeoning technology. Notable incidents, such as the Yahoo data breach in 2016, a DDOS attack on Microsoft's Skype in 2017, and the security flaw in Zoom reported by the Washington Post in 2020, underscore the substantial harm caused by data privacy and security issues in machine learning applications.

Currently, threats to machine learning security and privacy primarily fall into four categories: poisoning attacks [12] [13], adversarial sample attacks [14] [15], model extraction attacks [16], and model inversion attacks [17] see figure 1. Poisoning attacks and model inversion attacks occur during the training stage, where malicious data is injected to degrade model performance and information about the training set is obtained through reverse reasoning. Model extraction attacks and adversarial sample attacks take place during the inference phase, involving theft of internal model information and deception of the model by introducing interference factors to generate adversarial samples. Numerous defence measures have been developed to counter these threats, including homomorphic encryption [18], secure multi-party computation [19], and differential privacy [20].

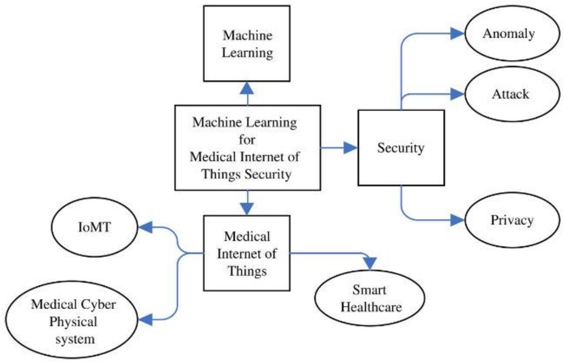

Figure 1: ML security and privacy approaches

The reliance on machine learning training on the quantity and quality of datasets poses a serious risk to widespread adoption due to the potential leakage of sensitive personal data. Model inversion attacks, particularly membership inference attacks, represent a critical privacy challenge by successfully inferring whether a specific target sample belongs to the target training dataset, resulting in privacy breaches. This attack has been successfully demonstrated in various data domains, such as biomedical data [21] [22] [23] and mobile location data [24], illustrating its potential harm to individual privacy and emphasizing the need for robust defence mechanisms.

Given that scholars specialize in various research fields with distinct problem-solving perspectives, the emphasis on member reasoning attack and defence varies among them. Thus, this paper initiates its exploration from the fundamental theory of attacking and defending machine learning models, scrutinizing pivotal technical models and elucidating the correlation between member inference attack models and the associated risks of privacy leakage. This endeavour holds immense significance in safeguarding data privacy and propelling advancements in the field of machine learning applications. The second section of this paper concisely outlines the adversary model, definition, classification, and generation mechanism of member inference attacks. In the subsequent sections, namely Sections 3 and 4, diverse types of member inference attack algorithms undergo detailed analysis, shedding light on their attack methods and current application status. Section 5 systematically organizes and summarizes the protective strategies employed against distinct attack methods, delving into the underlying reasons contributing to their effectiveness. Ultimately, Sections 6 and 7 encapsulate the comprehensive findings of the paper and present a forward-looking perspective for future research endeavours.

2. MEMBER INFERENCE ATTACK

In this section, we aim to consolidate and distill existing research findings on member inference attacks. Our focus is to succinctly summarize the key insights and methodologies explored in the current body of literature. This overview serves to provide a quick and informative reference for readers delving into the realm of member inference attacks.

2.1. Adversary Model

Within the domain of machine learning security, adversary models serve to delineate the capabilities and objectives of potential adversaries. In 2010, Barreno et al. [25] delved into the adversary model, considering both attacker capabilities and goals. Building upon this, Biggio et al. [26] expanded the adversary model in 2013 to encompass adversary goals, knowledge, capabilities, and strategies. The incorporation of these four dimensions offers a more systematic framework for characterizing the adversary's threat level when evaluating member reasoning

Table 1 Adversary model in membership inference attack

adversary model	describe
adversary target adversary knowledge	Breach of usability and privacy black box, white box
adversary capabilities adversary strategy	Strong adversary: can intervene in model training, access training data sets and collect intermediate results, etc.; Weak adversary: can only obtain model information or training data information through attack methods. Training phase: model reverse attack; Prediction stage: adversarial attack + member inference attack, model extraction attack + member inference attack

2.2. Definition and Model

Membership inference attacks involve the extraction of membership details from the training data by scrutinizing the target model system, constituting a prevalent type of attack leading to privacy breaches. This method determines whether specific data contributed to training the target model, enabling the attacker to infer details about the model's training set. As illustrated in Figure 2, the target model, trained on the original dataset, operates on the application platform. The attacker, posing as a user, accesses the target model, gathering relevant information and adversary knowledge to construct an attack model capable of deducing whether a given dataset constitutes a member of the training set.

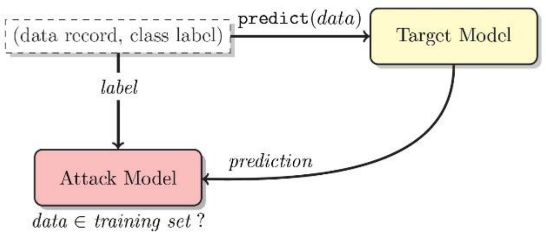

Figure 2: The model of membership inference attack

2.3. Categorization of Attack Models

Recent investigations into member inference attacks have resulted in categorizations based on distinct criteria, delineated in Table 2. Studies have classified these attacks into specific categories, each representing a unique standard or framework.

Table 2 Types of membership inference attacks

adversary knowledge	Attack method	attack mode	target model	type
black box	Shadow technology attack	passive aggressive	Classification model/deep learning/graph neural network/transfer learning	focus on learning
	baseline attack	passive aggressive	Classification model	focus on learning
	tag attack	passive aggressive	Classification model/deep learning	focus on learning
	diverted attack	passive aggressive	Classification model	focus on learning
White box	white box attack	Passive attack/active attack	Deep Learning/Generative Adversarial Networks	Centralized learning/federated learning

As indicated in Table 2, the classification of member inference attacks is based on the attacker's familiarity with the target model information, denoted as the adversary's knowledge. This results in two primary categories: black box attacks [27][33] and white box attacks [34][35]. In a black box attack, the attacker can solely access the model output results through the corresponding API, limited to observing the output $f(x;W)$ for input x without gaining access to intermediate results. Conversely, a white-box attack allows the attacker to access comprehensive information, including the target model's structure, training parameters, internal output results, training data distribution, and related data information.

Additionally, based on the attacker's engagement level, member inference attacks are further categorized into strong adversaries (active attacks) and weak adversaries (passive attacks). A strong adversary actively intervenes in the target model's training process, participating in federated learning and having the capability to modify intermediate data during training. In contrast, a weak adversary can only observe data changes during training and extract information through passive acquisition of the model interface.

Considering different attack types, member inference attacks primarily fall into two categories: centralized learning and federated learning. Centralized learning involves traditional model training with centralized storage of datasets for training the target model. On the other hand, federated learning entails local storage and training of personal data by each participant, exchanging gradients through a central parameter server for joint model training. Attackers in this model can either be a central parameter server or a local party.

Originally, member inference attacks predominantly targeted machine learning. However, with the widespread application of various data types such as images, text, and knowledge graphs, these attacks expanded to encompass transfer learning, deep learning, graph neural networks, and generative models. This broader scope has led to increased privacy risks.

2.4. Generating mechanism of attacks

The success of membership inference attacks hinges on a critical vulnerability known as overfitting within the target model. This susceptibility allows the model to memorize implicit traits of the training data, empowering attackers to discern membership relationships within the target data accurately. Additionally, factors like the introduction of abnormal data, characteristics of data distribution, and intermediate processes during model training furnish attackers with tools to detect targets and execute successful attacks.

Overfitting, a core component of membership inference attacks, involves attackers distinguishing between the training set and the test set of the target model. The model's proficiency in predicting the training set with high accuracy, coupled with diminished predictive abilities for the test set, renders models vulnerable to such attacks.

Outliers within the training set further exacerbate vulnerability. When these outliers, crucial for data representation, deviate in distribution from the test set data, the model's failure to adapt seamlessly results in distinguishability between the training set and test set. This distinctiveness facilitates the success of membership inference attacks.

Moreover, the impact of data and model factors, including shadow data set size, class and feature balance, and model configuration, contributes to the complexity of member inference attacks. These attacks are not solely influenced by one factor but rather orchestrated by the collaborative interplay of multiple factors.

3. ATTACK ALGORITHM

Membership inference attacks, demonstrated to be successful across diverse data domains, can be broadly categorized into two types within the realm of machine learning—those leveraging black-box knowledge and those reliant on white-box knowledge, as elaborated below.

3.1. Black box knowledge

The majority of studies on membership inference attacks have focused on black-box models. Shokri et al. were pioneers in proposing a membership inference attack on a machine learning model, successfully determining whether a specific patient had been discharged from the hospital [31]. Subsequently, Salem et al. introduced another attack by gradually relaxing Shokri et al.'s assumptions, achieving improved precision and recall [32]. Confidence-based membership inference attacks for machine learning models have also emerged in various domains, including federated learning, generative adversarial networks, natural language processing, transfer learning, and computer vision segmentation [33][38][39]. Decision-based attacks in the field involved Yeom et al.'s quantitative analysis of the relationship between attack performance and the loss of the training and test sets, introducing the first decision-based attack known as the Baseline attack [33]. Choo et al. proposed a method akin to boundary attack [38].

1. Shadow Technology Attack

The original membership inference attack against machine learning, known as the shadow technology attack, was proposed by Shokri. This approach necessitates the use of shadow technology to simulate the target model, constructing a training dataset to train the two-class attack model for membership inference [31]. As shown in Figure 3 .

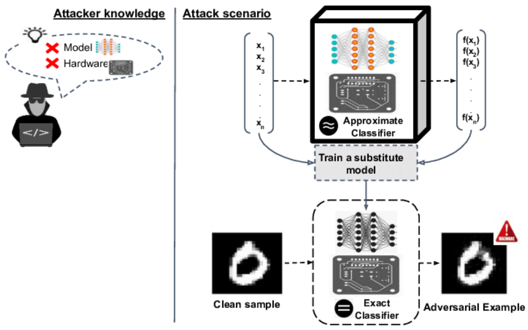

Figure 3: Black box attack

This methodology involves three primary steps: data synthesis, shadow model simulation, and attack model construction.

a) Data Synthesis: In situations where access to the model is restricted (black box scenario), the attacker lacks information about member data. Therefore, it becomes necessary to synthesize approximate data using various statistical algorithms such as model-based, statistical distribution-based, and noise-based methods.

b) Shadow Model Simulation: Relevant data synthesized in the previous step is employed to train one or more shadow models. These shadow models imitate the structure of the target model without having any knowledge about it. The shadow technology effectively simulates the target model through analysis and simulation, with the shadow model acting as a substitute for the original target model.

c) Attack Model Construction: Using the data set of the shadow model and the confidence vector output of the target model, a binary attack model is trained. This model, combined with the assigned label (where if data point x is lost to the training set of the shadow model, then label = 1; otherwise, label = -1), determines whether a given target data point belongs to the training data set of the target model.

Salem et al. [32] later relaxed Shokri's assumptions, proposing a more accurate and recall-focused approach. This method involves using only the output results of the target model for threshold discrimination, as shown in formula (1). While this approach is straightforward and highly efficient, its applicability is limited to models with poor generalization

Black-box attacks leveraging shadow technology initially focused on machine learning model API interfaces within cloud platforms, later expanding to include deep learning, transfer learning, and graph neural networks. In the context of shadowing attacks on graph neural networks trained on data like social networks and protein structures [34], synthetic data and shadow models may exhibit inconsistencies with the target system, yielding favourable outcomes even for models boasting strong generalization performance. This vulnerability in graph neural networks arises from heightened connectivity between instances.

2. Baseline Attack

Yeom et al. [33] introduced the baseline attack in 2018, performing membership inference based on the correct classification of data samples. If the target data is misclassified, it is deemed non-member data; otherwise, it is considered member data. The intensity of the baseline attack correlates positively with model overfitting. For models with substantial generalization gaps, the attack performance is high with low cost, but it proves ineffective for models exhibiting good generalization.

3. Tag Attack

Choo et al. [38] proposed a method resembling the boundary attack, conducted in a black-box setting solely with the target model's output label. This attack operates on the principle that training set samples are more resistant to perturbation than test set samples. The tag-based membership inference attack involves three stages:

a) Adversarial Sample Generation: Leveraging the target model's prediction label as input, adversarial sample technologies like FGSM, C&W, and hopskipjump induce decision changes on the target, generating adversarial samples.

b) Perturbation Mapping: Calculating the Euclidean distance between the adversarial sample and the original target, mapping the perturbation difficulty to distance categories to discern prediction differences between the target model's training and test data.

c) Member Inference: Logically distinguishing prediction differences to obtain fine-grained member signals for membership inference of the target group.

4. Diversion Attack

In [39], a diversion attack is proposed involving given data points (x, y) and the confidence vector obtained from the target model f(x). The cross-entropy loss loss(x, y) = − log(f(x)y) is calculated.

3.2. White Box Knowledge

In the realm of black-box knowledge attacks, the assailant is limited to targeting the training data solely based on the model's output. Nonetheless, the intermediate calculation data of the training process harbours substantial information about the training data. In pioneering work on attacking Generative Adversarial Networks (GANs), a white-box attack was first proposed, exclusively leveraging the output of the GAN's discriminator without learning the weights of the discriminator or generator to execute the attack. Furthermore, Nasr et al. extended the member inference attack to a white-box setting based on prior knowledge [35]. The activation function and gradient information obtained from the target model serve as inferred features for conducting member inference. The specific details are illustrated in Figure 4.

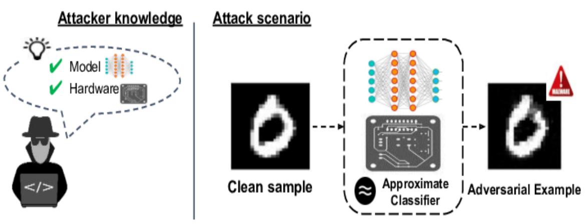

Figure 4: White box attack

Drawing from Figure 4, this solution operates on the principle that the target model undergoes fine-tuning and updates based on the training set data to minimize the loss gradient of the training data, thereby distinguishing between the gradient of the training set data and non-training set data for member inference.

For a target model f and input data x, the attacker computes the output of each layer in the forward propagation calculation of the target model, denoted as hi(x), the model output f(x), and the loss L(f(x; W), y). Subsequently, the gradient of each layer is calculated through backpropagation ∂L/∂Wi. These obtained parameters, along with the one-hot vector of y, constitute the input feature parameters of the attack model.

These input features are then fed into the corresponding Convolutional Neural Network (CNN) or Fully Connected Network (FCN) for feature extraction. The output is packaged and passed to the Fully Connected

Network (FCN), ultimately yielding the result of the inference attack. The attack model comprises two integral components: the Convolutional Neural Network (CNN) and the Fully Connected Network (FCN).

Additionally, Long et al. [37] introduced a member inference attack, GMIA, targeting well-generated models. In this attack, not all data is susceptible to member attacks. The attacker must identify vulnerable abnormal data points to differentiate members from non-members and execute a successful attack.

3.3. Algorithm comparison

In this section, we conduct a thorough comparison of the algorithms discussed earlier, providing an in-depth summary of existing member inference attack algorithms. The specific details of this comparative analysis are presented in Table 3.

Table 3 Comparison of membership inference attack algorithms

	adversary knowledge	target model	type	Assumptions			Attack accuracy	
				shadow model	Data distribution	Model structure	data set	accuracy(%)
[31]	black box	Classification model	independent model	yes	yes	yes	CIFAR100	92.8
[32]	black box	CNN	independent model	no	no	no	CIFAR100	85.7
[34]	White box	Classification model	federated learning	/	/	/	Yelp-health	75.0
[35]	White box	Classification model	federated learning	/	/	/	CIFAR100	85.1
[39]	black box	D.L.	independent model	no	yes	no	CIFAR10	88.0
[42]	black/white box	Generate model	independent model	no	no Yes	no	LFW	61.0/94.3
[43]	black box	NN	independent model	yes	no	no	Tweet(4)	64.8

4. CURRENT STATUS OF MEMBERSHIP INFERENCE ATTACKS

Given the ability of membership inference attacks to deduce the presence of specific data in a model's training set, their applications extend to verifying whether a user's data has been used without proper authorization. This capability has implications for disease monitoring, safety oversight, risk assessment, and privacy reinforcement in machine learning systems before potential attacks occur.

4.1. Auditing and Verification

Miao et al. [44] devised a voice audit model to identify if a user's voice data is part of the target model's training set, thereby indicating potential unauthorized use of user data. This user-centric member reasoning approach assesses whether a user's data was involuntarily utilized by the target model during training, promoting user rights protection and enabling audits of the target system model. Similarly, Song et al. [45] introduced an audit model for text generation models, deploying member inference to ascertain whether user data has been employed without proper authorization.

4.2. Disease Prediction

Membership inference attacks find application in disease monitoring using medical data [21] [22] [23] [36]. For instance, Homer et al. [21] aggregated profiles and case studies of target individuals with reference populations from public sources to determine if the target individual belongs to a group related to a particular disease. Moreover, in a diagnostic model developed from AIDS patient data, inferring that a person's medical data was used as the model's training data suggests a potential association with AIDS.

4.3. Safety Oversight and Intellectual Property Rights

Membership inference attacks prove useful in user credit monitoring [47] (e.g., one takeout platform serving multiple users), aggregate location monitoring [24], pre-release evaluation of privacy protection quality in systems (platforms), and regulatory authorities' monitoring for potential illegal use of user information, facilitating user rights protection. Additionally, these attacks pose a threat to the intellectual property rights of model providers over their training datasets.

5. DEFENCE STRATEGIES

In response to the diverse range of membership inference attacks, researchers have dedicated considerable attention to developing targeted defence solutions, leading to focused research efforts.

5.1. Defense Technologies

Member inference attacks pose a threat to the privacy of training set data. Defence strategies against membership inference fall into three main categories:

Regularization-Based Defenses [48] [49] [50]: These defences employ regularization techniques directly, including L2 regularization, dropout, model stacking, and min-max strategies.

Defence Based on Adversarial Attacks: This approach aims to protect the victim model through adversarial attacks.

Defence Based on Differential Privacy [51]: Differential privacy involves adding disturbance noise to various elements such as training data input, objective function, model gradient, and output processes to mitigate member privacy leakage.

The following outlines some of the latest defence technologies along with their advantages and disadvantages.

5.1.1. Min-Max Game

Nasr introduced a gaming concept to train models with membership privacy [48]. This approach ensures that the model remains indistinguishable between its training data and predictions for other data points. The privacy mechanism targets robust inference attacks, minimizing both privacy loss and classification loss. The optimization of the minimum-maximum objective function in this algorithm not only safeguards member privacy but also significantly mitigates the risk of overfitting.

5.1.2. mem-guard

mem-guard represents the inaugural defense mechanism that provides formal assurances regarding utility loss against membership inference [49]. Its core concept involves introducing carefully crafted noise to the confidence scores of the machine learning model, thereby misleading member classifiers. Essentially, the addition of a noise vector, denoted as "n," to the confidence score vector, "s," ensures a defense against membership inference attacks with guaranteed utility loss [41]. The algorithm seeks to identify the noise vector satisfying a unique utility-loss constraint.

Functioning as a defense against black box attacks, this algorithm probabilistically introduces noise to the confidence score vector obtained from the target model, forming a random noise addition mechanism. This allows the defender to simulate the attacker's attack classifier, creating a defense classifier, followed by the formulation of an optimization problem for resolution. Empirical evidence supports the assertion that mem-guard exhibits greater strength compared to min-max game and model stacking.

5.1.3. Differential Privacy

Chen's proposed differential privacy defense technology [51] safeguards model privacy by perturbing the model's weights. The mechanism entails a trade-off between privacy and model accuracy, where smaller privacy budgets offer more robust privacy guarantees at the expense of reduced model accuracy. Chen's experiments depict the relationship between the privacy budget and the accuracy of the target model as a logarithmic curve, identifying a balanced budget near the inflection point. Combining differential privacy with model sparsity substantially diminishes the vulnerability to membership inference attacks.

5.1.4. Other Defense Technologies

The MMD + Mix-up algorithm, introduced by Li [52], enhances the model's loss function by incorporating the maximum average difference between the softmax output empirical distributions of the training set and validation set as a regularizer. This regularization technique aims to minimize the distribution disparity between member and non-member samples, thereby fortifying the model against potential attacks.

6. CHALLENGESAND SUGGESTIONS

As artificial intelligence research and applications in machine learning continue to advance, the unique nature of machine learning algorithms presents substantial challenges for safeguarding user data and network models. Addressing these challenges requires a comprehensive consideration of heightened security and privacy threats, accompanied by the development of adaptable defence methods that enhance the efficacy of machine learning models. This section examines the research challenges associated with member inference attacks and defences, offering insights into future research directions.

Explore Efficient White-Box Knowledge-Based Machine Learning Member Inference Attacks

While current membership inference attacks based on black-box knowledge yield satisfactory performance across diverse datasets, their efficiency lags behind white-box attacks, imposing certain limitations. For instance, the efficacy of black-box shadow technology attacks is influenced by model generalization and constrained by assumptions regarding data distribution and model structure. Therefore, investigating efficient member inference attacks based on white-box knowledge becomes a pressing concern.

Develop a Generalized Membership Inference Attack Mechanism for Various Machine Learning Algorithms

Efforts are needed to design a membership inference attack mechanism that is universally applicable to different machine learning algorithms. Black-box attacks, primarily driven by overfitting, exhibit low efficiency and stability. Simultaneously, white-box attacks face coverage limitations in practical scenarios, particularly within federated learning contexts. A comprehensive approach that encompasses various machine learning algorithms and incorporates effective attribute inference is essential.

Devise Feasible Attack Plans for Non-Euclidean Spatial Data

Existing membership inference attacks predominantly focus on machine learning models trained on Euclidean space data, such as images and text. However, real-world data often manifests as graphs, as seen in social networks and protein structures. Current research has shown the viability of graph neural networks for processing such data, but privacy attacks on machine learning models in this realm remain underexplored. Exploring privacy preservation for non-Euclidean spaces without compromising the user experience in online social networks represents a promising avenue for research.

Strike a Balance Between Privacy, Efficiency, and Usability

Balancing the privacy of training data, model efficiency, and usability poses a significant challenge in machine learning. Privacy-preserving methods, such as differential privacy, may enhance privacy and efficiency but struggle to achieve an optimal utility-privacy balance due to added noise perturbation. Alternatively, secure multi-party computation offers high privacy and usability but introduces inefficiencies through noise perturbation and increased communication overhead. Establishing a multi-dimensional evaluation system and optimizing trade-offs among privacy, efficiency, and usability in diverse scenarios is crucial.

Establish a Unified Privacy Leakage Measurement Standard

In the realm of machine learning member inference attacks, measuring the privacy leakage risk of models is a critical aspect of evaluating attack performance. While some scholars have delved into privacy quantification, the research remains fragmented and narrowly focused on specific fields. A unified model and system for privacy leakage measurement and comprehensive risk analysis are yet to be established. Consequently, there is a need to develop a standardized privacy disclosure measurement and evaluation mechanism in machine learning.

Optimize Traditional Data Privacy Protection Solutions

Privacy protection solutions grounded in regularization, differential privacy, and adversarial games effectively mitigate privacy leakage in member inference attacks. However, given the sensitivity of private data and the model's robust memory capacity, there is room for optimization by combining traditional privacy defences with hybrid methods like cryptography, anonymity, adversarial regularization, and differential privacy. These optimizations can enhance overall data privacy protection.

7. CONCULOSION

This article initiates by presenting the current landscape of security and privacy threats confronting machine learning, delving into the intricacies of member inference attacks as part of the broader spectrum of data privacy threats. Subsequently, we conduct a comprehensive comparative analysis of prevalent member inference attack methods, exploring their application status. Following this, we scrutinize common privacy protection methodologies against member inference attacks and delve into the underlying mechanisms that render defense strategies successful. Ultimately, through an in-depth comparison and analysis of the limitations inherent in existing data privacy protection approaches, we address the challenges inherent in privacy protection research pertaining to member inference attacks, anticipating and preparing for more sophisticated attacks in the future.

REFERENCES

[1] Liu, Y., Ma, S., Aafer, Y., et al. (2018) Trojaning Attack on Neural Networks. Proceedings of the 25th Annual Network and Distributed System Security Symposium, San Diego, CA, 18-21 February 2018, 214-229. [DOI: 10.14722/ndss.2018.23291]

[2] Chen, S., Wang, H., Xu, F., et al. (2016) Target Classification Using the Deep Convolutional Networks for SAR Images. IEEE Transactions on Geoscience and Remote Sensing, 54, 4806-4817. [DOI: 10.1109/TGRS.2016.2551720]

[3] Fredrikson, M., Jha, S. and Ristenpart, T. (2015) Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. Proceedings of

the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, 12-16 October 2015, 1322-1333. [DOI: 10.1145/2810103.2813677]

[4] Jagannathan, G. and Wright, RN (2005) Privacy-Preserving Distributed k-Means Clustering over Arbitrarily Partitioned Data. Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Chicago, IL, 21-24 August 2005, 593-599. [DOI: 10.1145/1081870.1081942]

[5] Roy, A., Sun, J., Mahoney, R., et al. (2018) Deep Learning Detecting Fraud in Credit Card Transactions. 2018 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, 27 April 2018, 129-134. [DOI: 10.1109/SIEDS.2018.8374722]

[6] Jayaraman, B. and Evans, D. (2019) Evaluating Differentially Private Machine Learning in Practice. Proceedings of the 28th USENIX Conference on Security Symposium, Santa Clara, CA, 14-16 August 2019, 1895-1912.

[7] Liao Guohui, Liu Jiayong. Malicious code detection method based on data mining and machine learning [J]. Information Security Research, 2016, 2(1): 74-79.

[8] Tramèr, F., Zhang, F., Juels, A., et al. (2016) Stealing Machine Learning Models via Prediction APIs. Proceedings of the 25th USENIX Conference on Security Symposium, Austin, TX, 10-12 August 2016, 601-618.

[9] Gentry, C. (2009) Fully Homomorphic Encryption Using Ideal Lattices. Proceedings of the Forty-first Annual ACM Symposium on Theory of Computing, Bethesda, MD, 31 May 2009-2 June 2009, 169-178. [DOI: 10.1145/1536414.1536440]

[10] Chen, X., Xiang, S., Liu, CL, et al. (2014) Vehicle Detection in Satellite Images by Hybrid Deep Convolutional Neural Networks. IEEE Geoscience and Remote Sensing Letters, 11, 1797-1801. [DOI: 10.1109/LGRS.2014.2309695]

[11] Launchbury, J., Archer, D., DuBuisson, T., et al. (2014) Application-Scale Secure Multiparty Computation. In: Shao, Z., Ed., European Symposium on Programming Languages and Systems, Springer, Berlin, Heidelberg, 8-26. [DOI: 10.1007/978-3-642-54833-8_2]

[12] Han Ying, Li Shanshan, Chen Fuming. Seismic anomaly data mining model based on machine learning [J]. Computer Simulation, 2014, 31(11): 319-322.

[13] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., et al. (2017) Computer Aided Decision Making for Heart Disease Detection Using Hybrid Neural Network-Genetic Algorithm. Computer Methods and Programs in Biomedicine, 141, 19-26. [DOI: 10.1016/j.cmpb.2017.01.004]

[14] Fu, K., Cheng, D., Tu, Y., et al. (2016) Credit Card Fraud Detection Using Convolutional Neural Networks. In: Hirose, A., Ozawa, S., Doya, K., Ikeda, K., Lee, M. and Liu, D., Eds., International Conference on Neural Information Processing, Springer, Cham, 483-490. [DOI: 10.1007/978-3-319-46675-0_53]

[15] Jagielski, M., Oprea, A., Biggio, B., et al. (2018) Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. 2018 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, 20-24 May 2018, 19-35. [DOI: 10.1109/SP.2018.00057]

[16] Tian Chen. Application of evolutionary neural networks in credit card fraud detection[J]. Microelectronics and Computers, 2011, 28(10): 14-17.

[17] Acharya, UR, Oh, SL, Hagiwara, Y., et al. (2018) Deep Convolutional Neural Network for the Automated Detection and Diagnosis of Seizure Using EEG Signals. Computers in Biology and Medicine, 100, 270-278. [DOI: 10.1016/j.compbiomed.2017.09.017]

[18] Szegedy, C., Zaremba, W., Sutskever, I., et al. (2013) Intriguing Properties of Neural Networks. arXiv:1312.6199

[19] Papernot, N., McDaniel, P., Jha, S., et al. (2016) The Limitations of Deep Learning in Adversarial Settings. 2016 IEEE European Symposium on Security and Privacy (EuroS&P), Saarbruecken, 21-24 March 2016, 372-387. [DOI: 10.1109/EuroSP.2016.36]

[20] Jordan, MI and Mitchell, TM (2015) Machine Learning: Trends, Perspectives, and Prospects. Science, 349, 255-260. https://doi.org/10.1126/science.aaa8415

[21] Hagestedt, I., Zhang, Y., Humbert, M., et al. (2019) MBeacon: Privacy-Preserving Beacons for DNA Methylation Data. Proceedings of the 26th Annual Network and Distributed System Security Symposium, San Diego, CA, 24-27 February 2019, 72-87. [DOI: 10.14722/ndss.2019.23064]

[22] Li, J., Li, N. and Ribeiro, B. (2020) Membership

Inference Attacks and Defenses in Supervised Learning via Generalization Gap. arXiv:2002.12062

[23] Hui, B., Yang, Y., Yuan, H., et al. (2021) Practical Blind Membership Inference Attack via Differential Comparisons. arXiv:2101.01341. [DOI: 10.14722/ndss.2021.24293]

[24] Pyrgelis, A., Troncoso, C. and De Cristofaro, E. (2018) Knock Knock, Who's There? Membership Inference on Aggregate Location Data. Proceedings of the 25th Network and Distributed Systems Security Symposium, San Diego, CA, 18-21 February 2018, 199-213. [DOI: 10.14722/ndss.2018.23183]

[25] Yang, Z., Shao, B., Xuan, B., et al. (2020) Defending Model Inversion and Membership Inference Attacks via Prediction Purification. arXiv:2005.03915

[26] Barreno, M., Nelson, B., Joseph, AD, et al. (2010) The Security of Machine Learning. Machine Learning, 81, 121-148. [DOI: 10.1007/s10994-010-5188-5]

[27] Backes, M., Berrang, P., Humbert, M., et al. (2016) Membership Privacy in MicroRNA-Based Studies. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, Vienna, 24-28 October 2016, 319-330. [DOI: 10.1145/2976749.2978355]

[28] Homer, N., Szelinger, S., Redman, M., et al. (2008) Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. PLoS Genetics, 4, e1000167. [DOI: 10.1371/journal.pgen.1000167]

[29] Biggio, B., Fumera, G. and Roli, F. (2013) Security Evaluation of Pattern Classifiers under Attack. IEEE Transactions on Knowledge and Data Engineering, 26, 984-996. [DOI: 10.1109/TKDE.2013.57]

[30] Song, L., Shokri, R. and Mittal, P. (2019) Privacy Risks of Securing Machine Learning Models against Adversarial Examples. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, 11-15 November 2019, 241-257. [DOI: 10.1145/3319535.3354211].

[31] Melis, L., Song, C., De Cristofaro, E., et al. (2019) Exploiting Unintended Feature Leakage in Collaborative Learning. 2019 IEEE Symposium on Security and Privacy, San Francisco, CA, 19-23 May 2019, 691-706. [DOI: 10.1109/SP.2019.00029]

[32] Yeom, S., Giacomelli, I., Fredrikson, M., et al. (2018) Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. 2018 IEEE 31st Computer Security Foundations Symposium, Oxford, 9-12 July 2018, 268-282. [DOI: 10.1109/CSF.2018.00027]

[33] Wang, C., Liu, G., Huang, H., et al. (2019) MIASec: Enabling Data Indistinguishability against Membership Inference Attacks in MLaaS. IEEE Transactions on Sustainable Computing, 5, 365-376. [DOI: 10.1109/TSUSC.2019.2930526]

[34] Shokri, R., Stronati, M., Song, C., et al. (2017) Membership Inference Attacks against Machine Learning Models. 2017 IEEE Symposium on Security and Privacy, San Jose, CA, 22-26 May 2017, 3-18. [DOI: 10.1109/SP.2017.41]

[35] Yin, Y., Chen, K., Shou, L. and Chen, G. (2021) Defending Privacy Against More Knowledgeable Membership Inference Attackers. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14-18 August 2021, 2026-2036. [DOI: 10.1145/3447548.3467444]

[36] Nasr, M., Shokri, R. and Houmansadr, A. (2019) Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-Box Inference Attacks against Centralized and Federated Learning. 2019 IEEE Symposium on Security and Privacy, San Francisco, CA, 19-23 May 2019, 739-753. [DOI: 10.1109/SP.2019.00065]

[37] Long, Y., Bindschaedler, V., Wang, L., et al. (2018) Understanding Membership Inferences on Well-Generalized Learning Models. arXiv:1802.04889

[38] Choo, CAC, Tramer, F., Carlini, N., et al. (2020) Label-Only Membership Inference Attacks. arXiv:2007.14321

[39] Salem, A., Zhang, Y., Humbert, M., et al. (2019) ML-Leaks: Model and Data Independent [40] Membership Inference Attacks and Defenses on Machine Learning Models. Annual Network and Distributed System Security Symposium, San Diego, CA, 24-27 February 2019, 243-260. [DOI: 10.14722/ndss.2019.23119]

[41] Li, Z. and Zhang, Y. (2021) Membership Leakage in Label-Only Exposures. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, Korea, 15-19 November 2021, 880-895. [DOI: 10.1145/3460120.3484575].

[42] Hayes, J., Melis, L., Danezis, G. and De Cristofaro, E. (2019) LOGAN: Membership Inference Attacks against Generative Models. Proceedings on Privacy Enhancing Technologies, 2019, 133-152. [DOI: 10.2478/popets-2019-0008]

[43] Danhier, P., Massart, C. and Standaert, FX (2020) Fidelity Leakages: Applying Membership Inference Attacks to Preference Data. IEEE INFOCOM 2020-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, 6-9 July2020,728-733.[DOI:10.1109/INFOCOMWKSHPS50562.2020.9163032]

[44] Jia, J., Salem, A., Backes, M., et al. (2019) MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, 11-15 November 2019, 259-274. [DOI: 10.1145/3319535.3363201]

[45] Chen, J., Wang, WH and Shi, X. (2020) Differential Privacy Protection Against Membership Inference Attack on Machine Learning for Genomic Data. BIOCOMPUTING 2021: Proceedings of the Pacific Symposium, Kohala Coast, 3-7 January 2021, 26-37. [DOI: 10.1142/9789811232701_0003]

[46] Wang, Y., Wang, C., Wang, Z., et al. (2021) Against Membership Inference Attack: Pruning is All You Need. Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21), 3141-3147.

[47] Chen, J., Wang, WH, Gao, H., et al. (2021) PAR-GAN: Improving the Generalization of Generative Adversarial Networks against Membership Inference Attacks. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14-18 August 2021, 127-137.

[48] Hayes, J., Melis, L., Danezis, G. and De Cristofaro, E. (2019) LOGAN: Membership Inference Attacks against Generative Models. Proceedings on Privacy Enhancing Technologies, 2019, 133-152. [DOI: 10.2478/popets-2019-0008]

[49] Liu, G., Wang, C., Peng, K., et al. (2019) SocInf: Membership Inference Attacks on Social Media Health Data with Machine Learning. IEEE Transactions on Computational Social Systems, 6, 907-921. [DOI: 10.1109/TCSS.2019.2916086]

[50] Miao, Y., Zhao, BZH, Xue, M., et al. (2019) The Audio Auditor: Participant-Level Membership Inference in Voice-Based IoT. CCS Workshop of Privacy Preserving Machine Learning.

[51] Song, C. and Shmatikov, V. (2019) Auditing Data Provenance in Text-Generation Models. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, 4-8 August 2019, 196-206. [DOI: 10.1145/3292500.3330885]

[52] Fredrikson, M., Lantz, E., Jha, S., et al. (2014) Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. Proceedings of the 23rd USENIX conference on Security Symposium, San Diego , CA, 20-22 August 2014, 17-32.

[53] Nasr, M., Shokri, R. and Houmansadr, A. (2018) Machine Learning with Membership Privacy Using Adversarial Regularization. Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Toronto, 15-19 October 2018, 634-646. [DOI: 10.1145/3243734.3243855]

[54] Jia, J., Salem, A., Backes, M., et al. (2019) MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples. Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, 11-15 November 2019, 259-274. [DOI: 10.1145/3319535.3363201]

[55] Zheng, J., Cao, Y. and Wang, H. (2021) Resisting Membership Inference Attacks through Knowledge Distillation. Neurocomputing, 452, 114-126. [DOI: 10.1016/j.neucom.2021.04.082]

[56] Li, J., Li, N. and Ribeiro, B. (2021) Membership Inference Attacks and Defenses in Classification Models. Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy, USA, 26-28 April 2021, 5-16. [DOI: 10.1145/3422337.3447836]