

Institut für Informatik
Arbeitsgruppe ID Management
DER FREIEN UNIVERSITÄT BERLIN

Bachelorarbeit

**Taxonomy of Privacy Attacks
in Machine Learning**

Yarmina Anna Meszaros

anna.meszaros@fu-berlin.de

Betreuerin: M.Sc Franziska Boenisch
1. Gutachter: Prof. Dr. Marian Margraf
2. Gutachter(in):
Matrikel-Nr.: 4472877


Abgabetermin: 01. Dezember 2022

Eidesstattliche Erklärung

Ich versichere hiermit an Eides statt, dass diese Arbeit von niemand anderem als meiner Person verfasst worden ist. Alle verwendeten Hilfsmittel wie Berichte, Bücher, Internetseiten oder ähnliches sind im Literaturverzeichnis angegeben, Zitate aus fremden Arbeiten sind als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

30.11.22, Berlin

Datum, Ort



Unterschrift

Abstract

Machine learning has experienced a tremendous growth both in academic research as well as in real world applications. At the same time large amounts of data fuelled the concerns about security and privacy in machine learning. The result is a likewise growth of research in terms of privacy preserving machine learning.

To ensure unambiguous communication in academic research we need a clear understanding of what separates one threat from another and where they have similarities. Most existing taxonomies in this field are either too specialised or not specialised enough for our cause. Therefore, this work proposes a taxonomy which takes already existing ones into account and aims to offer a possible categorisation for future attacks. In order to do so this work focuses on five privacy attacks which are formally introduced and explained further.

Several existing papers on this topic were studied and compared to understand which are the most used and established terms and taxonomies in current literature. Besides that, the roles that play a part in private machine learning are presented and some possible criteria for classification into reasonable categories of privacy attacks are suggested. At the current time this is one of the first works that consider membership inference, model inversion, attribute inference, property inference and reconstruction attacks at the same time.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Problem Statement	1
1.3	Limitations	2
2	Background	2
2.1	Machine Learning	2
2.1.1	Types of Learning	2
2.1.2	Learning Architectures	3
2.1.3	Machine Learning as a Service	3
2.2	Roles	4
2.3	Adversary	5
3	Related Work	7
4	Attacks on Machine Learning	8
5	Privacy Attacks on Machine Learning	10
5.1	Membership Inference	11
5.2	Model Inversion	13
5.3	Attribute Inference	15
5.4	Property Inference	18
5.5	Reconstruction Attack	20
5.6	Relation between Privacy Attacks	22
6	Taxonomy	24
6.1	Existing Taxonomies	24
6.2	Criteria For Classification	25
6.3	Proposed Taxonomy	27
6.4	Terminology	27
6.5	Classification of Attacks	28
7	Conclusion	29
7.1	Results	29
7.2	Future Work	29
	References	35
A	Expanded Taxonomy	36
B	Expanded Classification	37

1 Introduction

1.1 Motivation

Machine learning has experienced a tremendous growth both in academic research as well as in real world applications. At the same time large amounts of data fuelled the concerns about security and privacy in machine learning. Our personal data are being used by several online services to train models which are used in machine learning applications. If models are trained on highly sensitive data like health records, it would be disastrous in terms of privacy if attackers would be able to extract specific information from a trained model about the used training data. Therefore, researching defence mechanisms and revealing security breaches is an increasingly important field. Essential for the collaborative understanding of different security threats and defence mechanisms are a uniform taxonomy and terminology. Whether we realise it or not, we are all inherent taxonomists. We classify things around us, because our daily life is filled with the need to distinguish between them and put the many objects around us into reasonable groups. And also, in academic research to ensure unambiguous communication and promote a successful development of countermeasures we need a clear understanding of what separates one threat from another and where they have similarities.

1.2 Problem Statement

In previous works some taxonomies on machine learning attacks were introduced [1, 2, 3, 4], but most of them only provide a partial coverage of privacy attacks or are too specialized to a narrow subset of attacks or cover machine learning in general. Additionally, most of them have many inconsistencies regarding the used terminology. (see Table 7) While some papers use certain terms synonymously other clearly differentiate them from each other. This thesis aims to propose a terminology which includes most of the state-of-the-art attacks on privacy in machine learning while taking existing taxonomies and terminologies into account and without eradicating existing definitions. Therefore, all previous used terms must be harmonised differentiated and set into relation. To do so this work aims to find similarities, differences and shows which criteria can be used to classify the attacks into different reasonable categories. This work focuses on five of the most prominent privacy attacks on machine learning: membership inference, model inversion, reconstruction attack, attribute inference and property inference and develops a taxonomy based on them.

34 papers on certain privacy attacks were studied and taken into account (see Tables 2, 3, 4, 5, 6) as well as several review paper to understand which are the most used and established terms and taxonomies. Especially recent survey papers will be considered to represent an as current as possible state of research.

1.3 Limitations

Despite the wide range of different attacks on machine learning this thesis will provide only a small introduction for attacks that do not target privacy and focus mainly on privacy attacks trying to propose a finer granulated division for this category. Even if this works tries to propose a timeless categorisation of machine learning privacy attacks it remains unclear if future attacks will fit clearly into the proposed taxonomy and if the categories will need to be updated or new ones will have to be introduced.

2 Background

2.1 Machine Learning

Machine learning (ML) is a fast-growing field devoted to the problem of learning from data without being explicitly programmed. In the following section a small overview over learning types and architectures will be given.

2.1.1 Types of Learning

Traditionally ML is split into three types of learning: supervised, unsupervised and reinforcement learning. New types and mixed types have emerged over the years.

Supervised Learning In a supervised learning setting, a model f with parameters θ represents the mapping function between inputs x and outputs $y = f(x, \theta)$, where x is a vector of attributes or features with dimensionality n : $x = (x_1, x_2, \dots, x_n)$. Nevertheless, the output or label y can have different dimensions. To train the model f_S a training data set S over the distribution \mathcal{D} is used, where $|S| = m$.

The mostly used supervised learning tasks are classification and regression. Examples for supervised learning algorithms include linear regression, logistic regression, decision trees or support vector machines but the majority of recent papers are focused on deep neural networks.

Unsupervised Learning The difference to supervised learning is, there are no labels y in unsupervised learning. The training set S only consists of the m inputs (x_1, \dots, x_n) . Unsupervised learning aims to identify structures or patterns in the data. Examples for unsupervised learning tasks are clustering, feature learning or anomaly detection.

Reinforcement Learning Reinforcement learning is concerned with intelligent agents that make observations of the environment and use these to take actions in order to maximize the notion of cumulative reward. Its focus is on finding a balance between exploration (of unknown territory) and exploitation (of current knowledge).

there are no privacy-related attacks against reinforcement learning at the current state, but it has been used to launch other privacy-related attacks [5].

Generative and Discriminative Learning Besides the three basic ML paradigms there are other categories like discriminative and generative algorithms.

Discriminative classifiers aim to model the conditional probability $P(y|x)$ in order to learn the decision boundaries that are separating the different classes based on the input data x . Algorithms like logistic regression and neural networks fall under this category.

Generative classifiers on the other hand try to model class and what are the features of the class. (Mathematically it tries to learn the joint probability distribution, $P(x,y)$, of the inputs x and label y .) When given a new observation, these classifiers try to predict which class would have most likely generated the given observation. An example of such classifiers is Naive Bayes. GANs which generate new data samples matching the properties of the training data also count as such classifiers.

2.1.2 Learning Architectures

The learning process is either a centralized or a distributed one. The main difference in that case is whether the data and the model are collocated or not.

Centralized Learning In centralized learning settings, the data and the model are collocated. This means the data used for the training of the model are gathered in one central place but can have multiple data producers or owners.

Distributed Learning Large amounts of data and memory capacity drive the need for distributed learning architectures. There are several variants of distributed learning like collaborative or federated learning, fully decentralized or peer-to-peer learning and split learning.

Collaborative or federated learning is the mostly common approach in recent literature. It has the goal to learn one global model from data stored in multiple remote devices or locations. The data do not leave the remote devices, are processed locally and then used to update the local models. These model updates are sent to the central server that aggregates them and creates a global model which the server then sends back to all participant devices.

2.1.3 Machine Learning as a Service

Machine Learning as a service (MLaaS) are cloud-based computing platforms that offer ML tools. Users are able to train and evaluate their models remotely or use pre-trained models which allows them to benefit from machine learning without the cost and time.

These tools are accessed via prediction APIs on a pay-per-query basis. Models from MLaaS normally are only available for input-output interaction without revealing the model architecture and parameters.

If a model is trained on a cloud-based server, it is possible its parameters and training hyperparameters may be revealed afterwards.

2.2 Roles

To understand who is threatened by which kind of attack it is useful to take a look at the participants, the assets and their roles and interests in the ML environment.

From a threat model perspective each asset that is potentially threatened (the training data set, the model itself, its parameters and its architecture) is linked to a participants who will be affected by an successful attack.

Rigaki et al. [5] identified following roles in this threat model:

- (1) **The data owners** whose data may be sensitive and need to be protected.
- (2) **The model owners**, which may not want to share information about their models.
- (3) **The model consumers** which use the services via user interfaces.
- (4) **The adversaries** which have access to the model's interfaces as a normal consumer does and if the model owner allows, they may have access to the model itself.

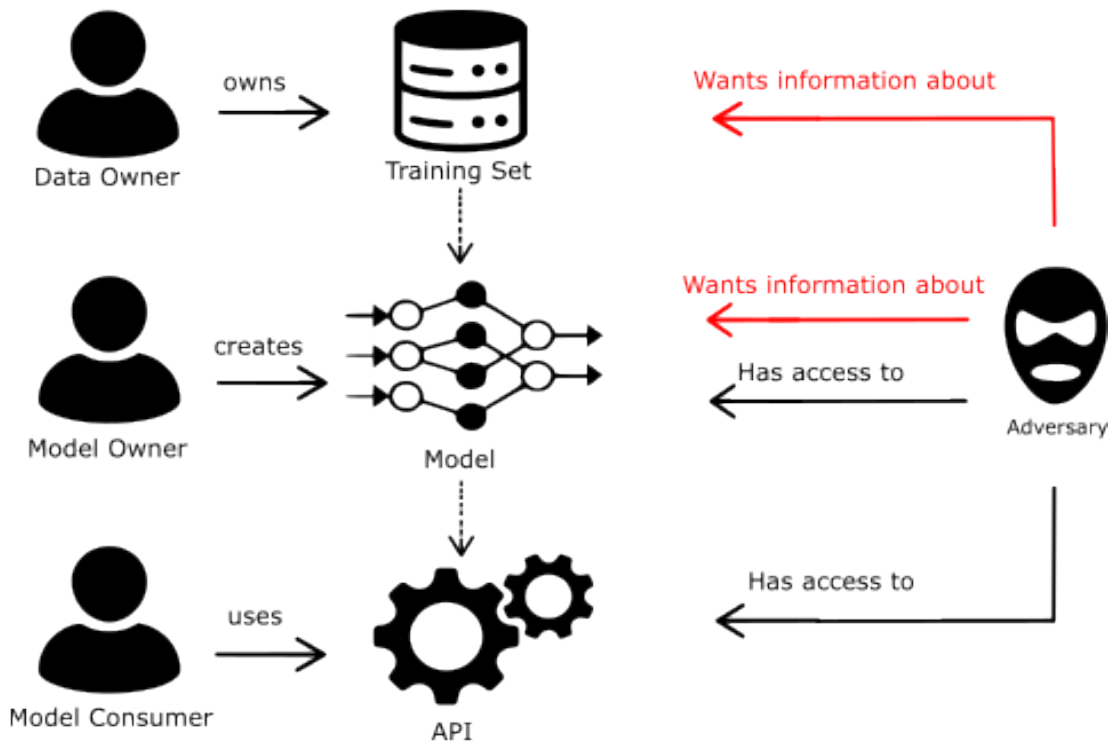


Figure 1: Threat Model of privacy and confidentiality attacks against machine learning systems.

2.3 Adversary

A role which needs special consideration is the adversary or attacker. In order to categorize and evaluate the attacks we need to elaborate the adversary’s capabilities: the knowledge about the target model and the data it was trained on (the original data), the adversary’s strategy and the target.

Based on the attacker’s capabilities, we can call the attacker weak or strong. For instance, an adversary that possesses the knowledge about the architecture of the target model and the original training data is stronger than an adversary that does not have that knowledge. Though it is not always possible to say which capabilities lead to a stronger attacker [4].

Adversary’s model knowledge

The adversary can receive different amounts of information of the target ML models. Depending on how much information the adversary has it can be easier to perform a certain attack [6].

Traditionally the possible settings are separated into black- and white-box access. Some papers also use the term grey-box access for situations in between [6] [7].

Black-box When talking about black-box access, also known as query access, we assume that the attacker only controls the input and has access to the output of the trained model. In most literature he can query the model as often as he needs to (he has unlimited resources). The adversary does not know the target model, including model architecture, model parameters, and training data. Therefore, the attacker needs to identify the model’s vulnerability only by utilizing knowledge about output responded from the target model.

White-box White-box access assumes that the user has full access to the trained model, including its input, output, architecture, and parameters. To identify the model’s vulnerability the adversary uses this available information to launch an attack. In some literature the adversary also has complete knowledge about the defence mechanisms that are used.

Grey-box Grey-box access describes situations which are in between. In some previous works it is characterised as a scenario where the adversary has access to the outputs of a model’s intermediate layers but has no knowledge about the architecture and parameters or vice versa [6].

In other literature the term grey-box is used for a specific situation, where the adversary has complete knowledge of the target model, including model architecture, model parameters, and training data and the only difference to a white-box setting is, that the adversary does not know the defence mechanism against the adversarial attack. The grey-box setting usually is used to evaluate the defence against the adversarial attack [7].

Adversary's data knowledge

Membership Describes the knowledge if an attacker knows a certain data point was part of the training data set. This knowledge can for example make it easier to perform a reconstruction attack [8].

Features Tells what the adversary knows about single features of some instances. For attribute inference attacks it is necessary to have knowledge about some non-sensitive features of a data point to be able to infer the sensitive ones.

Properties All the knowledge the attacker has about the data itself. The distribution, if it is evenly, how many class labels there are and everything that grants him an advantage.

Adversary's Strategy

We consider two different types of adversarial behaviour, passive and active. They are roughly reflecting the traditional distinction in security literature between honest-but-curious and fully malicious adversaries [9].

Passive The passive or honest-but-curious user interacts with the trained model only as it was intended by design. He is able to observe the updates and performs inference without changing anything in the training procedure.

All that can be revealed this way is involuntary leakage, if the model has any such vulnerability [6].

Active If an adversary interferes with the training in any way, they are considered an active attacker or malevolent user. But also, if the attacker tries to take advantage of potential vulnerabilities in the trained model, such as memorization and overfitting, aiming to extract sensitive data via privacy attacks.

Adversary's Target

In IT security we consider three main security goals: **Confidentiality**, **Integrity** and **Availability**. In context of machine learning they slightly differ from their traditional definition.

Additionally to the CIA Triad we consider privacy a protection target due to possibly endangered sensitive training data. Tabassi et al. [2] consider privacy violations a special subset of confidentiality violations.

Confidentiality Confidentiality is violated when an adversary extracts or infers usable information about the model or the data. This includes extraction attack that reveals model architecture or parameters, or an oracle attack that enables the adversary to construct a substitute model. But also attacks that reveal confidential information about the

data or a membership test to determine if an individual was included in the training data set of the target model.

Integrity With integrity violated, the process inference is undermined, which can lead to reduction of confidence or misclassification. In unsupervised learning this violation may produce a meaningless representation of the input in an unsupervised feature extractor while in Reinforcement Learning this may cause the learning agent to act unintelligently or with degraded performance.

Availability If the availability is violated this may induce reductions in quality (such as inference speed) or access (denial of service) to the point where the ML component is unavailable or unusable to users.

Privacy Privacy violations are considered a specific class of confidentiality violation where the data that is obtained is personal and sensitive information. For example, if an adversary extracts an individual's medical records in violation of privacy policies.

3 Related Work

One very detailed and comprehensive taxonomy for membership inference attacks in particular was proposed by Hu et al. [1]. This work presents a wide overview over performed attacks and all important parameters.

A taxonomy for model extraction attacks was introduced by Oliynyk et al. [4]. Therefore, several papers on that topic were observed. Special attention was given to the role of the adversary and his capabilities.

Tabassi et al. [2] developed a comprehensive taxonomy including all kinds of machine learning attacks, their defences and consequences. Additionally, it deals with the terminology used in machine learning security. All in all, this work offers a good overview of adversary machine learning in general.

A recent paper that addresses privacy in machine learning is the survey paper of Rigaki et al. [5]. It sheds light on the many aspects of machine learning privacy.

Liu et al. [10] established a threat model taxonomy focusing on four attacks – namely, membership inference, model inversion, attribute inference, and model stealing.

4 Attacks on Machine Learning

For all of these four security goals exist attacks which target them:

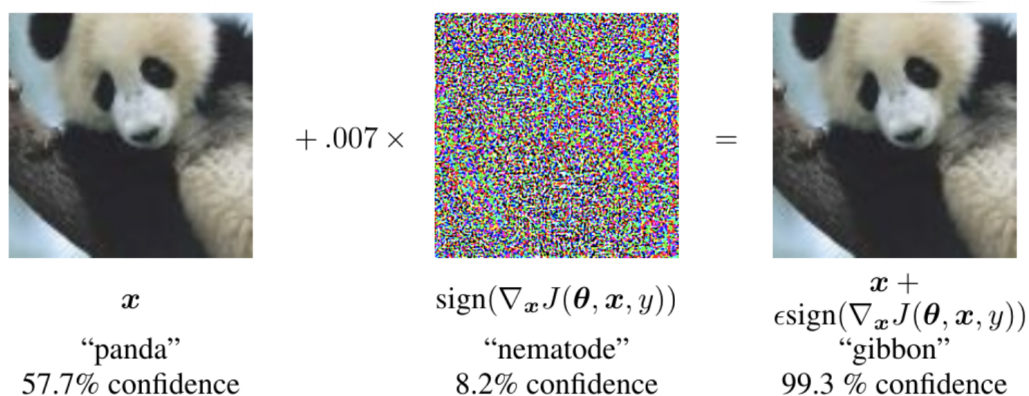
Security Goal	Attack	Timing
Integrity	Evasion Attacks	Inference Phase
Availability	Data Poisoning	Training Phase
Confidentiality	Model Stealing, Model Extraction	Inference Phase
Privacy	Model Inversion, Membership Inference, Attribute Inference, Property Inference	Inference Phase

Table 1: Attacks targeting security goals and their timing

In the following section the attacks that don't fall into the privacy category will be shortly introduced and looked at in regards of how they can facilitate a privacy attack. Model extraction and model for example often act as a stepping stone for privacy attacks [11, 12].

Evasion Attacks

Evasion attacks are some of the most common ones on machine learning models and are also referred to as adversarial examples. They are performed during inference. It refers to designing an input (adversarial example), which seems normal for a human but is wrongly classified by ML models. An often-used example is the alteration of some pixels in a picture, which leads the image recognition system to fail classifying the result correctly even if the human eye can't tell the difference.



(Goodfellow 2018)

Figure 2: A demonstration of fast adversarial example generation (Goodfellow et al., 2014 [13])

According to the goal of the adversary, the adversarial attack falls into two categories:

- **Non-targeted Attack**

The adversary crafts adversarial examples to cause the target model to misclassify the input with high confidence, but does not require the prediction to be specified class.

- **Targeted Attack**

The adversary crafts adversarial examples to cause the target model to misclassify the input with high confidence into a particular class t specified by the adversary.

Data Poisoning

Poisoning consists of contaminating the training data set such that the learner trains a bad classifier. This can lead the classifier to misclassify malicious sample or activities crafted by the adversary at the testing stage. This way malicious samples, modify data labels, and corruption could be injected into the training data.

A particular case of data poisoning is called backdoor attack, which aims to teach a specific behaviour for inputs with a given trigger, e.g., a small defect on images, sounds, videos or texts. In [14] data poisoning was used as a steppingstone for a model inversion attack.

Data poisoning has also been used to increase the amount of sensitive information a model leaks about a particular sensitive attribute [6].

Depending on the attacker's goals, the poisoning attack falls into three categories:

- **Accuracy Drop Attack**

The adversary aims to disrupt the training process by injecting malicious samples to reduce the performance of the target model at the testing stage.

- **Target Misclassification Attack**

The adversary aims to enforce test samples to be misclassified at the testing stage.

- **Backdoor Attack**

The adversary aims to install a backdoor with a specific mark so that the target model has a target output for that particular input.

Model Stealing

Besides personal data machine learning systems have a second valuable part - the model itself. This has value because of the amount of work that goes into the development of well performing models.

So firstly, the intellectual property of the model owners is targeted and secondly extracting the complete model gains the attacker white box access, which makes many attacks much more powerful [15]. Often model stealing attacks are seen as a stepping stone for further attacks, (e.g. membership inference [12]) which is why they are often mentioned in context of privacy attacks especially in survey papers.

Model Extraction

Model extraction is a special form of model stealing in which an adversary aims to steal parameters of the target model. This is done with a black-box access to the target model and extracting information. Goal is to extract a (nearly) equivalent substitute model which behaves very similarly to the model under attack.

Like privacy attacks the model extraction falls under the category of so-called oracle attacks [2] which query a trained model to extract useful information.

Functionality Extraction

Rather than stealing the model itself, here the goal is to create “knock-offs” of the (black-box) model solely based on input-output pairs observed from MLaaS queries. Even if these models are just knock-offs still the intellectual property is violated [9].

5 Privacy Attacks on Machine Learning

Private Data

To better understand the role of different kinds of data in machine learning security we distinguish between personal, personal “sensitive” and non-personal data.

Personal data is defined in the Article 4 of the GDPR and refers to data that directly or indirectly relates to an identified or identifiable natural person.

Sensitive data is defined by the GDPR as the personal data revealing racial or ethnic origin, political opinions, religious beliefs, health-related data or data concerning a person’s sex life or sexual orientation. This data is particularly worthy of protection and data processing, storage, transfer has to happen with special care.

Non-personal data Everything outside the scope of personal is non-personal data. This does not necessarily mean that it’s leakage can be seen uncritical. Data that has to personal connection at first sight may help to recover sensitive data in some cases.

5.1 Membership Inference

Goal

Membership Inference Attacks (MIA) are a very popular category of privacy attacks in machine learning. The attacker's goal is to determine whether a data point was part of the training set which was used to train the target model.

Motivation

Even though this may not at first seem to pose a serious privacy risk, sensitive data could be potentially revealed. For example when it is determined if an individual's data was used to train a model, which is trained to predict certain health conditions one can assume this individual suffers from this health condition.

Definition

In a MIA, the adversary attempts to infer whether a specific point was included in the data set used to train a given model.

The adversary is given a data point $z = (x; y)$, access to a model $A(S)$, the size of the model's training set $|S| = n$, and the distribution \mathcal{D} that the training set was drawn from. With this information the adversary must decide whether $z \in S$.

Experiment 1 (Membership experiment $\text{Exp}^{\text{MIA}}(\mathcal{A}; A; n; \mathcal{D})$). Let \mathcal{A} be an adversary, A be a learning algorithm, n be a positive integer, and \mathcal{D} be a distribution over data points $(x; y)$. The membership experiment proceeds as follows:

1. Sample $S \sim \mathcal{D}^n$, and let $A_S = A(S)$.
2. Choose $b \leftarrow \{0; 1\}$ uniformly at random.
3. Draw $z \in S$ if $b = 0$, or $z \notin S$ if $b = 1$
4. $\text{Exp}^{\text{MIA}}(\mathcal{A}; A; n; \mathcal{D})$ is 1 if $\mathcal{A}(z; A_S; n; \mathcal{D}) = b$ and 0 otherwise. \mathcal{A} must output either 0 or 1.

Definition 1 (Membership advantage). The membership advantage of \mathcal{A} is defined as

$$\text{Adv}^{\text{MIA}}(\mathcal{A}; A; n; \mathcal{D}) = \Pr[\text{Exp}^{\text{MIA}}(\mathcal{A}; A; n; \mathcal{D}) = 1 | b = 1] - \Pr[\text{Exp}^{\text{MIA}}(\mathcal{A}; A; n; \mathcal{D}) = 1 | b = 0]$$

where the probabilities are taken over the coin flips of \mathcal{A} , the random choices of S and b , and the random data point $z \in S$ or $z \notin S$.

Early work

Shokri et al. [11] first introduced this kind of attack. They demonstrated an attack, where the adversary, uses black-box queries to a supervised ML model, trying to verify whether certain data records were used to train the model. To accomplish that attacker trains a some so called shadow models which are imitating the target model's behaviour, where the number of shadow models are distributed similarly to the target model's training data. Using this shadow models the adversary finally trains an attack model which distinguishes whether a sample was used in the training of the target or not.

Further attacks

Salem et al. [16] performed a quite successful membership inference attacks with only one shadow model.

Nasr et al. [17] redesigned the attack by using one-hot encoded class labels as part of input features and training a single NN attack classifier for all class labels.

Causes

One well known cause that was shown to improve the accuracy of MIA in black-box attacks is poor generalisation of the model [11]. The effect of overfitting was later confirmed by Yeom et al. [18].

Additionally Truex et al. [19] shed light on some causes and circumstances that influence the efficiency of a MIA. The data seems to matter a lot. The more classes the target model was trained on the more vulnerable it was, because each class takes up a smaller region in the latent space and there is less uninformed space. the algorithm also plays an important role. Algorithms whose decision boundaries are unlikely to be drastically impacted by a particular instance are more resilient.

Limitations

MIAs are not suitable for every kind of machine learning task. In order to create an attack model that works efficiently, the adversary must be able to explore the feature space, So if we have to deal with a complex image classification for high resolution photos the cost of creating training examples for the attack will not be reasonable anymore.

Reference		Access		Algorithm
Author	Year	Black-box	White-box	
Shokri et al. [11]	2017	•		Neural network
Hayes et al. [20]	2018	•	•	GAN
Long et al. [21]	2018	•		Neural network
Melis et al. [22]	2018		•	Neural network
Salem et al. [16]	2018	•		Neural network
Yeom et al. [18]	2018	•		Neural network, DT, Linear Regression
Nasr et al. [17]	2018	•		Neural network
Nasr et al. [12]	2019		•	Neural network
Chen et al. [23]	2020	•	•	GAN

Table 2: Observed Membership Inference Attacks

5.2 Model Inversion

Goal

Model Inversion (MI) attacks are a type of privacy attack that tries to recover average representation of the training classes given access only to a trained classifier. This attack specifically aims to extract an average representation of each of the classes the model was trained on given that the adversary has access to a model (either black-box or white-box). This can allow an adversary to reconstruct average representation of training classes which were included in the training data without previous knowledge of it. Though, they're unable to extract individual instances.

Motivation

A reconstruction of training data can lead to a violation of the individual's privacy. If we regard the software Faception, which is marketed by an Israeli concern with the same name [24]. Faception is a machine learning model which draws conclusions from an individual's face on the personality. The developers state that their model is able to determine if a portrait belongs to "an Extrovert, a person with High IQ, Professional Poker Player or a Terrorist" [24]. Especially in view of the latter category an adversary could have a special interest in knowing how an person of this category would look like on average.

Definition

In MI the adversary aims to infer an average representation of a given class. The attacker is given a class label y and has to reconstruct an average representation of the m features (x_1, \dots, x_m) . We assume this representation is given by $Ave(x)$.

Model inversion is formalized in Experiment 2 where the adversary is given a class label y .

Experiment 2 (Model inversion experiment $\text{Exp}^{\text{MI}}(\mathcal{A}; A; n; \mathcal{D})$). Let \mathcal{A} be an adversary, n be a positive integer, and \mathcal{D} be a distribution over data points $(x; y)$. The model inversion experiment proceeds as follows:

1. Sample $S \sim \mathcal{D}^n$, and let $A_S = A(S)$.
2. Choose $b \leftarrow \{0; 1\}$ uniformly at random.
3. Draw $z \in S$ if $b = 0$, or $z \notin S$ if $b = 1$
4. $\text{Exp}^{\text{MI}}(\mathcal{A}; A; n; \mathcal{D})$ is 1 if $\mathcal{A}(\varphi(z); A(S); n; \mathcal{D}) = \text{Ave}(x_i)$ and 0 otherwise.

Definition 2 (Model inversion advantage). The model inversion advantage of \mathcal{A} is defined as

$$\text{Adv}^{\text{MI}}(\mathcal{A}; A; n; \mathcal{D}) = \Pr[\text{Exp}^{\text{MI}}(\mathcal{A}; A; n; \mathcal{D}) = 1 | b = 1] - \Pr[\text{Exp}^{\text{MI}}(\mathcal{A}; A; n; \mathcal{D}) = 1 | b = 0]$$

where the probabilities are taken over the coin flips of \mathcal{A} , the random choices of S and b , and the random data point $z \in S$ or $z \notin S$.

Early work

The concept of MI is introduced by Fredrikson et al. [25]. They succeeded with reconstructing recognizable characteristics of a portrait from training data of a facial recognition model. The attacker was given only API access to a facial recognition system and the name of the person whose face is recognized by it. In order to do so they used so-called “hill climbing”. They repeated targeted queries which they slightly adjusted every time to maximize the output probabilities step by step.

In academic literature this example can be often found, however this case is an extreme one, because every output class of the model represents an individual person. This means every data point belonging to a class are just different pictures of the same person. So, if the adversary tries to create an input which leads to a high confidence score of a target class he just gets a mean value of all pictures belonging to the same class, not an actual single data point.

Terminology

The term model inversion is in some cases used as an umbrella term for all kinds of reconstruction attacks like attribute inference and in rare cases even property inference [26]. In [27] they can most likely be counted to class-wise representation reconstruction attacks. Rigaki et al. [5] vice versa counted it with attribute inference as part of a group of reconstruction attacks.

This work refers to model inversion exclusively for attacks that reveal an average representation of each of the classes.

Further attacks

Hitaj et al. [28] were able to construct class representatives using GANs. They also used models where all members of the same class are visually similar (handwritten digits and faces).

Hidano et al. [14] managed to perform an attack where no knowledge of nonsensitive features was needed. They made their attack work by poisoning training data properly.

Causes

Yeom et al. [18] showed that a higher generalization error can lead to a higher probability to infer data attributes, but also that the influence of the target feature on the model is an important factor.

Limitations

Attacks like this only seem to be successful if the if the data point of one class does not contain too wide variation. In this case all the training data points which belong to one class were frontal pictures of the same person and the attacker was able to reconstruct the mean value of them. As soon as we have many varying pictures, the reconstructions will be unusable because the object won't be recognisable anymore. This was demonstrated by Shokri et al. [11].

Reference		Access		Algorithm
Author	Year	Black-box	White-box	
Fredrikson et al. [25]	2015	•	•	DT, Neural network
Wu et al. [29]	2016	•	•	Logistic Regression, Neural network
Hitaj et al. [28]	2017		•	Neural network
Hidano et al. [14]	2017		•	Linear Regression
Zhao et al. [30]	2022	•		Neural Network

Table 3: Observed Model Inversion Attacks

5.3 Attribute Inference

Goal

Attribute Inference (AI) attacks are often referred to as model inversion. But other than in classical model inversion attacks [25] here the adversary is able to actual reconstruct features from single data points. In this kind of attack an adversary is given access to a model as well as incomplete information about a data point. This can be some non-sensitive attributes. The goal is to infer the missing information for that point, just as sensitive attributes.

Motivation

Non-sensitive data is probably not as protected as sensitive. Still it can be used to run this attack. It should not be possible for an attacker who already knows the name, age and weight of training data point to infer other sensitive data like blood type or medical conditions.

Definition

We assume that data points are triples $z = (v; t; y)$, where $(v; t) = x \in X$ and t is the sensitive features targeted in the attack. The fixed function φ describes the information about data points known by the adversary. Let T be the support of t when $z = (v; t; y) \in \mathcal{D}$. The function π is the projection of X into T where $\pi(z) = t$.

Attribute inference is formalized in Experiment 3 where the adversary is given partial information $\varphi(z)$ about the challenge point z .

Experiment 3 (Attribute Inference experiment $\text{Exp}^{\text{AI}}(\mathcal{A}; A; n; \mathcal{D})$). *Let \mathcal{A} be an adversary, n be a positive integer, and \mathcal{D} be a distribution over data points $(x; y)$. The attribute experiment proceeds as follows:*

1. *Sample $S \sim \mathcal{D}^n$, and let $A_S = A(S)$.*
2. *Choose $b \leftarrow \{0; 1\}$ uniformly at random.*
3. *Draw $z \in S$ if $b = 0$, or $z \notin S$ if $b = 1$*
4. *$\text{Exp}^{\text{AI}}(\mathcal{A}; A; n; \mathcal{D})$ is 1 if $\mathcal{A}(\varphi(z); A(S); n; \mathcal{D}) = \pi(z)$ and 0 otherwise.*

Definition 3 (Attribute advantage). *The Attribute advantage of \mathcal{A} is defined as*

$$\text{Adv}^{\text{AI}}(\mathcal{A}; A; n; \mathcal{D}) = \Pr[\text{Exp}^{\text{AI}}(\mathcal{A}; A; n; \mathcal{D}) = 1 | b = 1] - \Pr[\text{Exp}^{\text{AI}}(\mathcal{A}; A; n; \mathcal{D}) = 1 | b = 0]$$

where the probabilities are taken over the coin flips of \mathcal{A} , the random choices of S and b , and the random data point $z \in S$ or $z \notin S$.

Early work

Fredrikson et al. [31] first proposed this AI attack, which is related to the model inversion attack. By using black-box access and partial information about an individual's medical record the attacker was able to recover the individual's genomic information. Given a linear regression model that predicted a suggested dose of the drug Warfarin using a feature vector consisting of patient demographic information and medical history they were able to infer the sensitive attribute: the genetic markers. Their algorithm completes the target feature vector with each of the possible values and then computes a weighted probability estimate that this is the correct value.

Terminology

In literature the term attribute inference has also been used in a different way to describe attacks where an attacker uses publicly accessible data to infer sensitive "attributes" [32, 33, 34]. This kind of attacks will not be included in this taxonomy because they target the individual's data directly instead of training data of an ML model.

Other literature counts attribute inference to model inversion [5, 35]. This thesis differentiates between these two attacks and refers to attacks that actually reconstructs sensitive attributes of a data point instead of a representation over the classes in the training data.

Further Attacks

Yeom et al. [36] examined the connections between AI and MIAs and showed a reduction from the membership inference attack to the attribute inference attack and vice versa.

Causes

It is suggested that AI, like MIA, is indeed sensitive to overfitting [36]. Regardless of the generalization error, the attacker's ability to learn more about the training data also increases the chance of a successful attack.

Limitation

At least some non-sensitive attributes are previously needed to perform that attack and infer the targeted sensitive attribute.

Reference		Access		Algorithm
Author	Year	Black-box	White-box	
Fredrikson et al. [31]	2014		•	Linear Regression
Yeom et al. [18]	2018	•		Linear Regression, DT, Neural network
Wang et al. [37]	2018		•	Neural network
Zhang et al. [38]	2020		•	Neural network
Mehnaz et al. [39]	2020	•		DT, Neural network
Yeom et al. [36]	2020	•		Linear Regression, DT, Neural network
Zhao et al. [40]	2021	•	•	Neural network

Table 4: Observed Attribute Inference Attacks

5.4 Property Inference

Goal

A Property Inference attack (PIA) aims to extract information about the data set properties which are not correlated to the learning task. This can for example be the ratio of men and women in a medical data set where the sex was not an encoded attribute.

Motivation

PIAs can lead to draw conclusions about the distribution of the training data like gender or age. This can be problematic in some scenarios. For example, if a business trains a model with their client's data. An attack could reveal the demography of their customer base.

Definition

In a PIA, the adversary aims to infer training data set properties of a given model. This property is unrelated to the main classification task of the model.

The adversary has access to a model $A(S)$, the size of the model's training set $|S| = n$, and the distribution \mathcal{D} that the training set was drawn from. He has also a property P about the data set which he to classify as true or false.

Experiment 4 (Property experiment $\text{Exp}^{\text{PIA}}(\mathcal{A}; A; n; \mathcal{D})$). *Let \mathcal{A} be an adversary, n be a positive integer, $P(S)$ a property regarding the training data set S and \mathcal{D} be a distribution over data points $(x; y)$. The property experiment proceeds as follows:*

1. Sample $S \sim \mathcal{D}^n$, and let $A_S = A(S)$.
2. Choose $b \leftarrow \{0; 1\}$ uniformly at random.
3. Draw $P(S) = \text{false}$ if $b = 0$, or $P(S) = \text{true}$ if $b = 1$
4. $\text{Exp}^{\text{PIA}}(\mathcal{A}; A; n; \mathcal{D})$ is 1 if $\mathcal{A}(z; A(S); n; \mathcal{D}) = b$ and 0 otherwise. \mathcal{A} must output either 0 or 1.

Definition 4 (Property advantage). *The property advantage of \mathcal{A} is defined as*

$$\text{Adv}^{\text{PIA}}(\mathcal{A}; A; n; \mathcal{D}) = \Pr[\text{Exp}^{\text{PIA}}(\mathcal{A}; A; n; \mathcal{D}) = 1 | b = 1] - \Pr[\text{Exp}^{\text{PIA}}(\mathcal{A}; A; n; \mathcal{D}) = 1 | b = 0]$$

where the probabilities are taken over the coin flips of \mathcal{A} , the random choices of S and b , and the random data point $z \in S$ or $z \notin S$.

Early work

Ateniese et al. [41] are one of the first, to extract “something meaningful relating to properties of the training set. They investigated if a speech recognition classifier was trained only with people who speak an Indian-English dialect.

Further attacks

Ganju et al. [42] transitioned this approach to fully connected networks where the weights and biases of the neural networks were used as input to the meta-classifier.

Wang et al. [43] proposed three kinds of PIAs: class sniffing, quantity inference, and whole determination. Class sniffing detects whether a training label is present within a training round. Quantity inference determines how many clients have a given training label in their data set. The whole determination infers the global proportion of a specific label.

Causes

PIAs are possible even with well-generalized models [42, 22] so overfitting does not seem to be a cause of PIAs. Unfortunately, regarding property inference attacks, we have less information about what makes them possible and under which circumstances they appear to be effective.

Limitations

To train the meta-classifier, the attacker first needs training data to train the shadow classifier. This might not be the case in practice. However, there are many existing approaches to facilitate the generation of training data. For example, the attacker could generate synthetic training data for the shadow models using model inversion attacks

Reference		Access		
Author	Year	Black-box	White-box	Algorithm
Ateniese et al. [41]	2015	•		SVM, HMM
Melis et al. [22]	2018		•	Neural network
Ganju et al. [42]	2018		•	Neural network
Wang et al. [43]	2019	•		Neural network
Parisot et al. [26]	2021	•		Neural network
Zhou et al. [44]	2021	•		GAN

Table 5: Observed Property Inference Attacks

5.5 Reconstruction Attack

Goal

The goal of a reconstruction attack (RA) is to reconstruct test data partially or fully. This is often done by external knowledge on feature vectors or the data used to build the ML model. These feature vectors can be used to reconstruct the raw data. This usually requires white-box access to the ML models deployed.

Another approach is to use deep leakage from gradients [45, 46].

Motivation

Reconstruction of sensitive data out of publicly available data means a huge privacy violation. Especially if the recovered information can be used to hack into further systems and even deal more damage to an individual's privacy.

Definition

In a RA, the adversary aims to infer raw training data points of a given model. The adversary has access to a model $A(S)$, the size of the model's training set $|S| = n$, and the distribution \mathcal{D} that the training set was drawn from and feature vectors $\vec{f} = (f_1, \dots, f_m)$ corresponding to a data point z (in some cases gradients).

His goal is to reconstruct raw data $r = \text{rec}(z)$.

Experiment 5 (Reconstruction experiment $\text{Exp}^{\text{RA}}(\mathcal{A}; A; n; \mathcal{D})$). *Let \mathcal{A} be an adversary, n be a positive integer and \mathcal{D} be a distribution over data points $z = (x; y)$. The reconstruction experiment proceeds as follows:*

1. Sample $S \sim \mathcal{D}^n$, and let $A_S = A(S)$.
2. Choose $b \leftarrow \{0; 1\}$ uniformly at random.
3. Draw $z \in S$ if $b = 0$, or $z \notin S$ if $b = 1$.
4. $\text{Exp}^{\text{RA}}(\mathcal{A}; A; n; \mathcal{D})$ is 1 if $\mathcal{A}(z; A(S); n; \mathcal{D}) = \text{rec}(z)$ and 0 otherwise.

Definition 5 (Reconstruction advantage). *The reconstruction advantage of \mathcal{A} is defined as*

$$\text{Adv}^{\text{RA}}(\mathcal{A}; A; n; \mathcal{D}) = \Pr[\text{Exp}^{\text{RA}}(\mathcal{A}; A; n; \mathcal{D}) = 1 | b = 1] - \Pr[\text{Exp}^{\text{RA}}(\mathcal{A}; A; n; \mathcal{D}) = 1 | b = 0]$$

where the probabilities are taken over the coin flips of \mathcal{A} , the random choices of S and b , and the random data point $z \in S$ or $z \notin S$.

Early work

Feng et al. [47] demonstrated in 2011 an attack to reconstruct fingerprint images directly from minutiae templates. The main idea was to reconstruct the phase image from minutiae which is then converted into the original grayscale image and then to launch an attack against fingerprint recognition systems to infer private data.

Terminology

The term reconstruction attack is sometimes mistakenly used as an umbrella term for all kind of attacks that are reconstruction training data or gain information about it. In other words all attacks that target *attribute privacy* [5]. Some literature refers to database reconstruction attacks on public data [48]. In regards of ML most refer to the reconstruction of raw data from feature vectors ([49]) or more recently to leakage from gradients ([45, 46, 27])

Further attacks

Al-Rubaie et al. [49] researched RAs using gestures raw data from user's authentication profiles. They were able to utilize the actual feature vectors that were stored in the user profiles to reconstruct raw data which then to use that information to hack into other systems.

Deep Leakage from Gradient (DLG) [45] was the first work to fully reveal the private training data from gradients, which can obtain the training inputs as well as the labels in only a few iterations. This approach was improved by Zhao et al. in 2020 [46].

Causes

These attacks are especially possible if the feature vectors used during the training phase to build the ML model were not flushed from the server after the model was finished. Also ML models that store explicit feature vectors (e.g. SVM) should be avoided.

Limitations

In both DLG [45] and iDLG [46] are several weaknesses that may limit their applicabilities. Both adopted a computationally expensive second-order optimization method, both are only suitable for the gradient computed on a small batch of samples and both works used untrained model neglecting gradients over multiple communication rounds. [8]

Reference		Access		
Author	Year	Black-box	White-box	Algorithm
Al-Rubaie et al. [50]	2016		•	SVM
Oh et al. [51]	2019		•	Neural networks
Zhu et al. [45]	2019		•	Neural network, GAN
Zhao et al. [46]	2020		•	Neural network, GAN
Lyu et al. [8]	2021		•	Neural Network
Xie et al. [52]	2021		•	Neural Network
Balle et al. [53]	2022		•	Neural Network

Table 6: Observed Reconstruction Attacks

5.6 Relation between Privacy Attacks

Model Inversion - Attribute Inference

Although they either often used synonymously or summarized under reconstruction attacks there are differences. They are both related because both of them use access to the model to reconstruct some of the training data. But whereas the model inversion attack only reveals an average representation of each of the classes (see Fig. 3), the attribute inference can actually reconstruct some features of data points, which is a different threat.

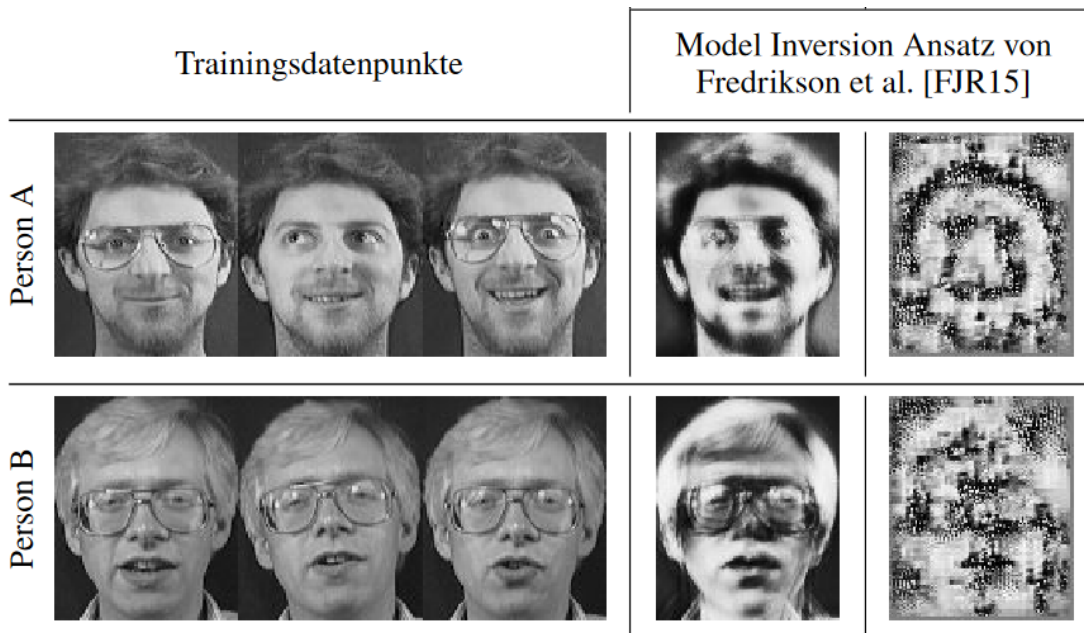


Figure 3: Model Inversion reconstructs the average representation - not a single image (Battis et al., 2021 [54])

Membership Inference - Attribute Inference

Yeom et al. [36] researched the connections between membership inference and attribute inference attacks and suggest that attribute inference may be more difficult than membership inference attribute because advantage implies membership advantage. They showed that these two attacks are closely related through reductions in both directions. Further their results confirm that models become more vulnerable to both types of attacks as they overfit more.

Model Inversion - Reconstruction Attack

While both attacks aim to infer some of the training data, the most important difference is that reconstruction attacks infer single data points while model inversion is only able to reproduce an average representation.

Sometimes these two terms are also often used for one another (see Table 7). Other literature uses reconstruction as an umbrella term for model inversion and attribute inference attacks [5].

Al-Rubaie et al. [49] however proposes a differentiation between model inversion and reconstruction, stating that model inversion is only able to restore feature vectors while reconstruction manages to recover actual raw data.

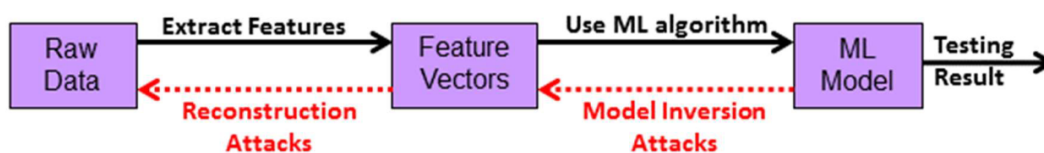


Figure 4: Model Inversion and Reconstruction Attack in comparison (Al-Rubaie et al., 2018)

Attribute Inference - Reconstruction Attack

Both attacks aim to infer or reconstruct attributes of data points. In attribute inference attacks the adversary has in the most cases access to some attributes of a data point and tries to recover the sensitive attributes correlation to given data point.

Whereas in reconstruction attacks the attacker often has access to the feature vectors and aims to reconstruct raw data.

6 Taxonomy

6.1 Existing Taxonomies

Most already existing taxonomies are either too specialised or not specialised enough for our cause. But the following shown taxonomies were taken into account when creating one for privacy attacks following their example.

Hu et al. [1] developed a very detailed taxonomy regarding membership inference attacks and assigned conducted attacks to the categories. For our matter this taxonomy is too specialised to membership inference attacks and therefore not suitable for privacy attacks in general. However, this approach could be suitable if a finer granulated classification is desired and could be considered in future work.

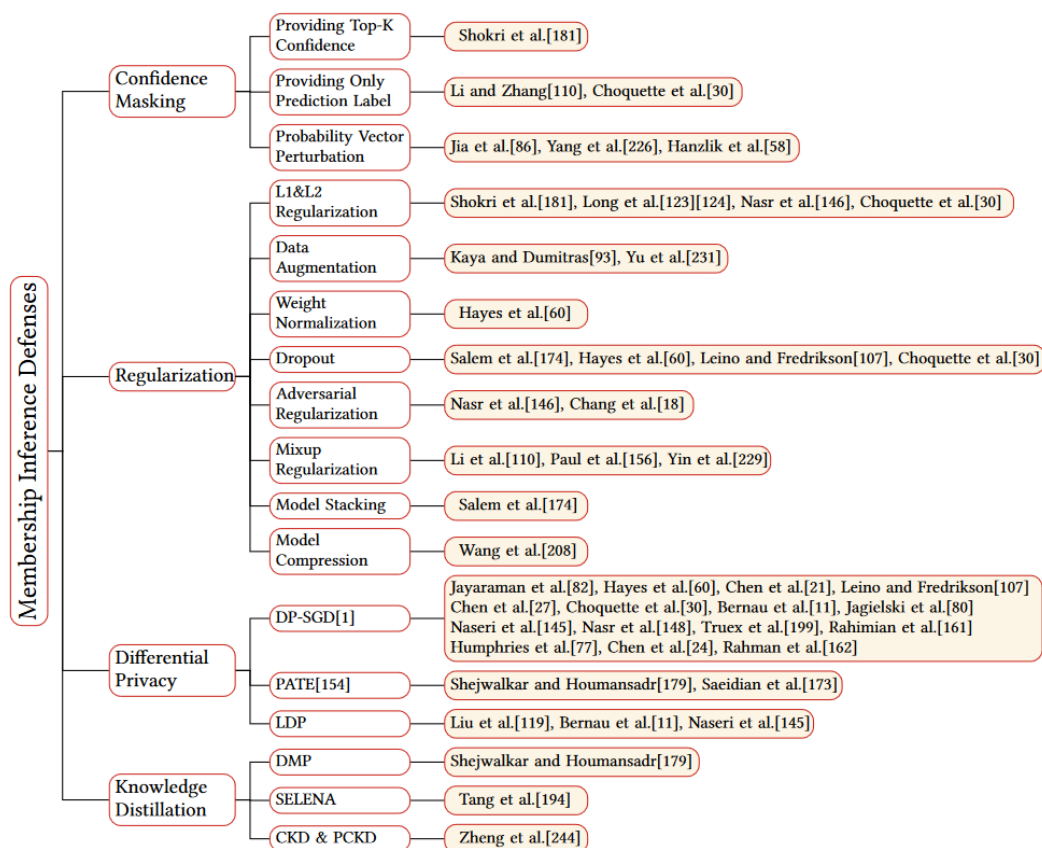


Figure 5: Taxonomy of Model Extraction attacks (Hu et al., 2022)

Oliynyk et al. [4] however introduced a less fine granulated taxonomy tailored to model extraction attacks. The kind of attacks do not suit the privacy-related attacks, but the degree of detail seemed just about right for the causes of this work.

Pitropakis et al. [3] suggest a taxonomy for adversarial attacks on machine learning in general based on the attacker’s knowledge and capabilities.

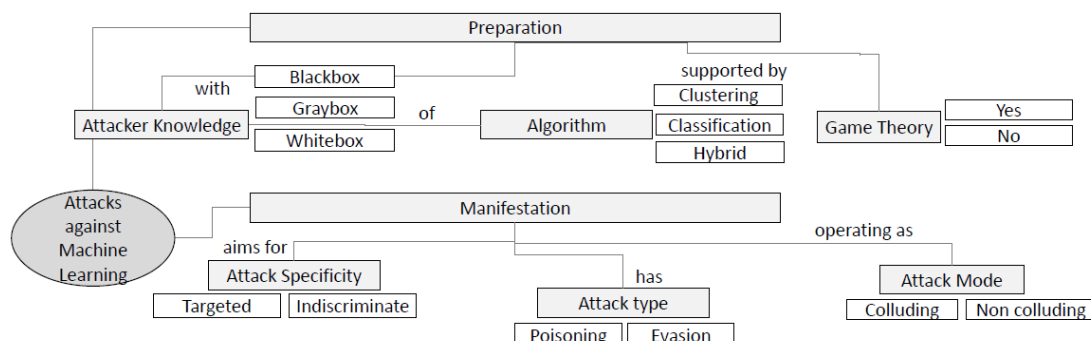


Figure 6: Taxonomy of Adversarial attacks (Pitropakis et al., 2019)

Tabassi et al. [2] created a taxonomy for adversarial machine learning regarding attacks, defences and consequences. This taxonomy seems to be the most general but detailed taxonomy at that time. But both are too general regarding our focus on privacy attacks.

6.2 Criteria For Classification

Timing of the attack

One variable is where the attack takes place. This can happen at two times: *the training phase* or *the inference phase*.

- **Training Phase:** the adversary attempts to manipulate the training data which will be used to train the model. This can be done by injecting data or changing labels for example.
- **Inference Phase:** the adversary collects information about the trained model by querying it and observing inferences made by it

Timing of attack correlates in most cases with the target. If we know the targeted security goals and the appropriate attack, we know the time the attack has to be performed. An overview over which attacks target which security goal is shown in Table 1.

All our privacy attacks take place while inference that's why in a privacy attack focused taxonomy the timing may seem to be irrelevant but previous works have shown that poisoning the data beforehand may simplify following privacy attacks [14]. Therefore, it could be interesting to know if an attacker had the opportunity to manipulate training data.

Targeted Security Goal

In the previous section it was shown that the targeted security goals in most cases sets the timing of the attack. An overview and explanation over the security goals was given in

subsection 2.3. Our attacks revolve around privacy, so we want to take a look at different kinds of privacy. One useful subdivision of privacy is the distinction between *membership privacy* and *attribute privacy* [55].

- **Membership privacy** refers to the information if a particular data point was included into the training set and therefore is a member of it. Depending on the kind of ML model this can lead to violation of privacy. For example, if an individual was member in a data set that contains people with a specific health condition knowledge about the membership this individual reveals status about this health condition.
- **Attribute privacy** is about protecting potentially sensitive attributes of individual data points. An attacker who already knows about a data point that is member of the training data set should not be able to infer more sensitive attributes about.

While membership inference attacks target the membership privacy, model inversion and attribute inference attack the attribute privacy.

Another subdivision that can be made is between the *individual's privacy* and the *group privacy*. While model inversion, membership inference and attribute inference target the privacy of an individual, property inference mostly just reveals information about the distribution over the whole training data set.

Target

Another factor that needs to be considered is the kind of target of an attack. Our four suggested privacy attacks all aim to infer something of the used training data, Model extraction attacks on the contrary aim to steal the model or its functionality. In the proposed taxonomy model extraction attacks will be taken into account because stealing or extracting the model will possibly "open" the model for the adversary so that he will be able to perform following attacks with a white-box access.

Kind of data

The kind of data that is used to train the model plays also a role in regards of suitable privacy attacks. While in Fredrikson et al. [25] model inversion attack it was easy to recognise an individual on the average presentation of the used frontal pictures, the mean value of some data with categorical values may be not be as useful for an attacker. Furthermore, membership inference attacks are not easy to perform with complex image classification for high resolution photos.

Adversary's capabilities

As already mentioned in Subsection 2.3 adversaries can have a variety of capabilities regarding knowledge and strategy. This factor has to be taken into account when looking at attacks because different access to the target model can result in a different of success.

6.3 Proposed Taxonomy

Regarding all these factors the following taxonomy was developed:

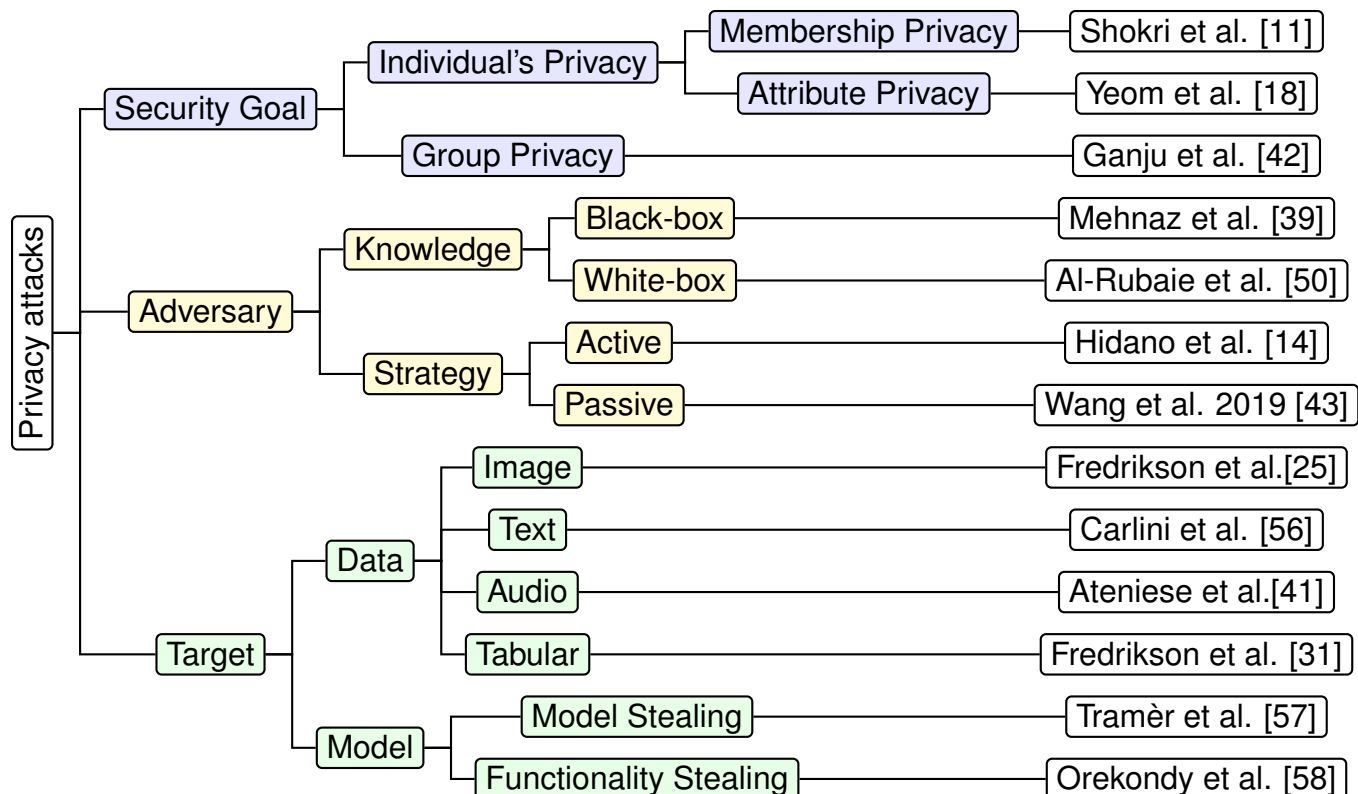


Figure 7: Proposed Taxonomy with chosen examples.

6.4 Terminology

As stated in Section 5 the used terms in literature are often misleading and inconsistent. Therefore an overview is introduced to clarify the usage of used terms in this work compared to the used terms in other works.

Terminology used	Used synonymously for	Used as a generic term for
Membership Inference	linkage attacks [6] tracing [59][17][60][61]	
Model Inversion	Input Inference, data extraction[15], reconstruction [6][5], attribute inference[5][35]	membership inference [7][26], reconstruction attack, Property inference [26]
Attribute Inference	reconstruction[5], model inversion[5][35]	
Model Extraction	Parameter Inference[15], Model stealing[10][4]	Property Inference [6]
Reconstruction	Attribute inference [5] [6]	Attribute Inference, Model Inversion[5]

Table 7: The first column gives the term primarily used within literature. The second and third column lists other terms used across literature

6.5 Classification of Attacks

Regarding our introduced terminology and the criteria for classification of machine learning attacks a classification system for existing privacy attacks is proposed which aims to allow future attacks to be classified within this system. The main criteria which were considered were the attack timing and the targeted security goals where our different kinds of privacy were considered.

For non-privacy attacks this categorisation can be easily refined and expanded in a similar way. An approach for that is given in appendix B.

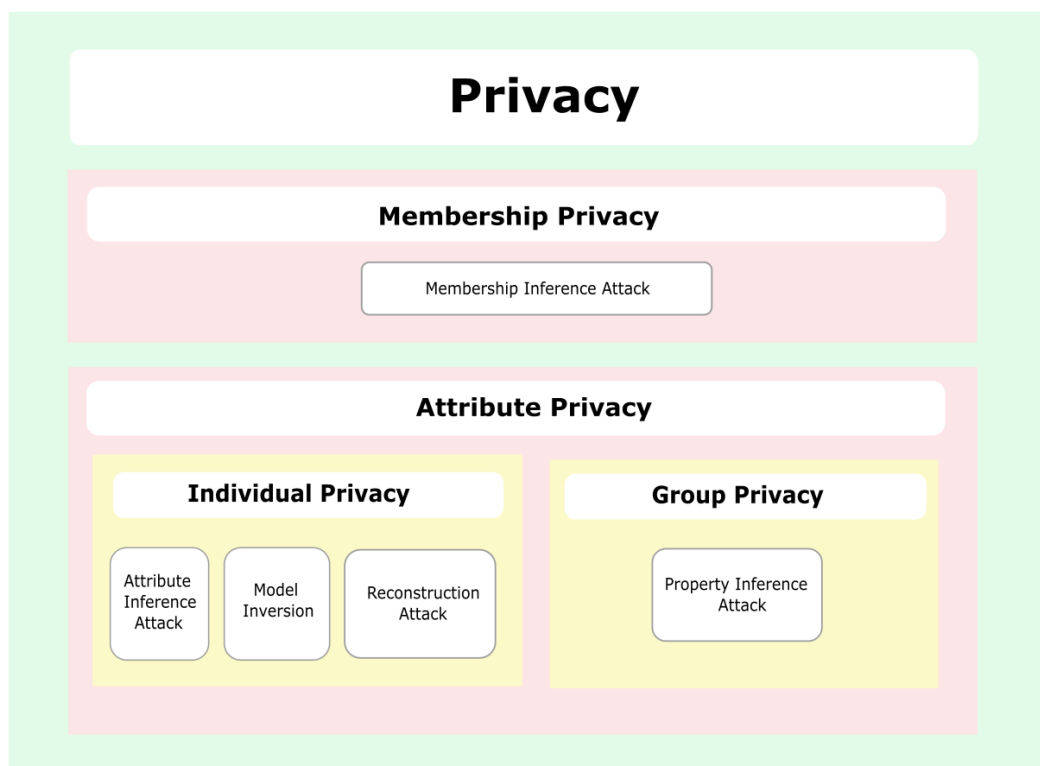


Figure 8: Classification of privacy attacks

7 Conclusion

7.1 Results

Five privacy attacks were introduced and distinguished between. But also, similarities and relations between them were discussed and shown. It was pointed out that in literature some inconsistencies regarding the terminology of machine learning attacks exist and a general terminology was proposed that was orientated towards the already most common used one. Based on chosen survey paper and our five attacks a taxonomy and categorisation were introduced that allows any to this date existing attacks to be classified. Exemplary some specific attacks were chosen and categorised. The proposed taxonomy does not contradict any of the already existing ones and tries to take every definition and subset of attacks into account.

7.2 Future Work

Especially in term of terminology there needs to be some clarification to prevent misunderstandings. In particular the terms reconstruction attacks as well as attribute inference and model inversion would benefit some clear definition.

Additionally it could be a future task to extend the proposed taxonomy and categorisation. This could comprise providing a refined and expanded description of the evasion and poisoning attacks which are only considered superficial in this work. As new findings occur in research taxonomy and classification can be expanded or refined for new machine learning attacks. A suggestion how this could be approached is shown in Appendix A and B.

References

- [1] H. Hu, Z. Salcic, G. Dobbie, and X. Zhang, “Membership inference attacks on machine learning: A survey,” *CoRR*, vol. abs/2103.07853, 2021.
- [2] E. Tabassi, K. Burns, M. Hadjimichael, A. Molina-Markham, and J. Sexton, “A taxonomy and terminology of adversarial machine learning,” 10 2019.
- [3] N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas, “A taxonomy and survey of attacks against machine learning,” *Computer Science Review*, vol. 34, October 2019.
- [4] D. Oliynyk, R. Mayer, and A. Rauber, “I know what you trained last summer: A survey on stealing machine learning models and defences,” 2022.
- [5] M. Rigaki and S. Garcia, “A survey of privacy attacks in machine learning,” 2020.
- [6] M. Jegorova, C. Kaul, C. Mayor, A. Q. O’Neil, A. Weir, R. Murray-Smith, and S. A. Tsafaris, “Survey: Leakage and privacy at inference time,” *CoRR*, vol. abs/2107.01614, 2021.
- [7] X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, and A. V. Vasilakos, “Privacy and security issues in deep learning: A survey,” *IEEE Access*, vol. 9, pp. 4566–4593, 2021.
- [8] L. Lyu and C. Chen, “A novel attribute reconstruction attack in federated learning,” *CoRR*, vol. abs/2108.06910, 2021.
- [9] E. D. Cristofaro, “An overview of privacy in machine learning,” *ArXiv*, vol. abs/2005.08679, 2020.
- [10] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. De Cristofaro, M. Fritz, and Y. Zhang, “MI-doctor: Holistic risk assessment of inference attacks against machine learning models,” 02 2021.
- [11] R. Shokri, M. Stronati, and V. Shmatikov, “Membership inference attacks against machine learning models,” *CoRR*, vol. abs/1610.05820, 2016.
- [12] M. Nasr, R. Shokri, and A. Houmansadr, “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)*, IEEE, may 2019.
- [13] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” 2014.

-
- [14] S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, and G. Hanaoka, “Model inversion attacks for prediction systems: Without knowledge of non-sensitive attributes,” in *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, pp. 115–11509, 2017.
- [15] A. Polyakov, “How to attack Machine Learning (Evasion, Poisoning, Inference, Trojans, Backdoors).” <https://towardsdatascience.com/how-to-attack-machine-learning-evasion-poisoning-inference-trojans-back> 2019. [Online; accessed 19-July-2022].
- [16] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, “MI-leaks: Model and data independent membership inference attacks and defenses on machine learning models,” 2018.
- [17] M. Nasr, R. Shokri, and A. Houmansadr, “Machine learning with membership privacy using adversarial regularization,” 2018.
- [18] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, “Privacy risk in machine learning: Analyzing the connection to overfitting,” 2017.
- [19] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, “Demystifying membership inference attacks in machine learning as a service,” *IEEE Transactions on Services Computing*, vol. 14, no. 6, pp. 2073–2089, 2021.
- [20] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro, “LOGAN: evaluating privacy leakage of generative models using generative adversarial networks,” *CoRR*, vol. abs/1705.07663, 2017.
- [21] Y. Long, V. Bindschaedler, L. Wang, D. Bu, X. Wang, H. Tang, C. A. Gunter, and K. Chen, “Understanding membership inferences on well-generalized learning models,” *CoRR*, vol. abs/1802.04889, 2018.
- [22] L. Melis, C. Song, E. D. Cristofaro, and V. Shmatikov, “Inference attacks against collaborative learning,” *CoRR*, vol. abs/1805.04049, 2018.
- [23] D. Chen, N. Yu, Y. Zhang, and M. Fritz, “Gan-leaks: A taxonomy of membership inference attacks against gans,” *CoRR*, vol. abs/1909.03935, 2019.
- [24] “Faception Facial Personality Analysis.” <https://www.faception.com/>. [Online; accessed 29-July-2022].
- [25] M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, CCS ’15*, (New York, NY, USA), p. 1322–1333, Association for Computing Machinery, 2015.

-
- [26] M. P. M. Parisot, B. Pejo, and D. Spagnuolo, “Property inference attacks on convolutional neural networks: Influence and implications of target model’s complexity,” 2021.
- [27] F. Boenisch, A. Dziedzic, R. Schuster, A. S. Shamsabadi, I. Shumailov, and N. Papernot, “When the curious abandon honesty: Federated learning is not private,” *CoRR*, vol. abs/2112.02918, 2021.
- [28] B. Hitaj, G. Ateniese, and F. Pérez-Cruz, “Deep models under the GAN: information leakage from collaborative deep learning,” *CoRR*, vol. abs/1702.07464, 2017.
- [29] X. Wu, M. Fredrikson, S. Jha, and J. F. Naughton, “A methodology for formalizing model-inversion attacks,” in *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, pp. 355–370, 2016.
- [30] X. Zhao, W. Zhang, X. Xiao, and B. Y. Lim, “Exploiting explanations for model inversion attacks,” 2022.
- [31] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, “Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing,” *Proceedings of the ... USENIX Security Symposium. UNIX Security Symposium*, vol. 2014, pp. 17–32, 08 2014.
- [32] J. Jia and N. Z. Gong, “Attriguard: A practical defense against attribute inference attacks via adversarial machine learning,” *CoRR*, vol. abs/1805.04810, 2018.
- [33] N. Gong and B. Liu, “You are who you know and how you behave: Attribute inference attacks via users’ social friends and behaviors,” 06 2016.
- [34] Y. Piao, K. Ye, and X. Cui, “Privacy inference attack against users in online social networks: A literature review,” *IEEE Access*, vol. 9, pp. 40417–40431, 2021.
- [35] A. Wainakh, E. Zimmer, S. Subedi, J. Keim, T. Grube, S. Karuppayah, A. S. Guinea, and M. Mühlhäuser, “Federated learning attacks revisited: A critical discussion of gaps, assumptions, and evaluation setups,” *CoRR*, vol. abs/2111.03363, 2021.
- [36] S. Yeom, I. Giacomelli, A. Menaged, M. Fredrikson, and S. Jha, “Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning,” *J. Comput. Secur.*, vol. 28, p. 35–70, jan 2020.
- [37] K.-C. Wang, Y. Fu, K. Li, A. Khisti, R. Zemel, and A. Makhzani, “Variational model inversion attacks,” 2022.
- [38] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, “The secret revealer: Generative model-inversion attacks against deep neural networks,” 2019.

-
- [39] S. Mehnaz, N. Li, and E. Bertino, “Black-box model inversion attribute inference attacks on classification models,” 2020.
- [40] B. Z. H. Zhao, A. Agrawal, C. Coburn, H. J. Asghar, R. Bhaskar, M. A. Kaafar, D. Webb, and P. Dickinson, “On the (in)feasibility of attribute inference attacks on machine learning models,” in *2021 IEEE European Symposium on Security and Privacy (EuroSP)*, pp. 232–251, 2021.
- [41] G. Ateniese, G. Felici, L. V. Mancini, A. Spognardi, A. Villani, and D. Vitali, “Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers,” *CoRR*, vol. abs/1306.4447, 2013.
- [42] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, “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, CCS ’18*, (New York, NY, USA), p. 619–633, Association for Computing Machinery, 2018.
- [43] L. Wang, S. Xu, X. Wang, and Q. Zhu, “Eavesdrop the composition proportion of training labels in federated learning,” 2019.
- [44] J. Zhou, Y. Chen, C. Shen, and Y. Zhang, “Property inference attacks against gans,” 11 2021.
- [45] L. Zhu, Z. Liu, and S. Han, “Deep leakage from gradients,” *CoRR*, vol. abs/1906.08935, 2019.
- [46] B. Zhao, K. R. Mopuri, and H. Bilen, “idl: Improved deep leakage from gradients,” *CoRR*, vol. abs/2001.02610, 2020.
- [47] J. Feng and A. K. Jain, “Fingerprint reconstruction: From minutiae to phase,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 2, pp. 209–223, 2011.
- [48] S. Garfinkel, J. Abowd, and C. Martindale, “Understanding database reconstruction attacks on public data: These attacks on statistical databases are no longer a theoretical danger,” *Queue*, vol. 16, 09 2018.
- [49] M. Al-Rubaie and J. M. Chang, “Privacy-preserving machine learning: Threats and solutions,” *IEEE Security Privacy*, vol. 17, no. 2, pp. 49–58, 2019.
- [50] M. Al-Rubaie and J. M. Chang, “Reconstruction attacks against mobile-based continuous authentication systems in the cloud,” *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 12, pp. 2648–2663, 2016.

-
- [51] H. Oh and Y. Lee, “Exploring image reconstruction attack in deep learning computation offloading,” in *The 3rd International Workshop on Deep Learning for Mobile Systems and Applications*, EMDL ’19, (New York, NY, USA), p. 19–24, Association for Computing Machinery, 2019.
- [52] S. Xie and Y. Hong, “Reconstruction attack on instance encoding for language understanding,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, (Online and Punta Cana, Dominican Republic), pp. 2038–2044, Association for Computational Linguistics, Nov. 2021.
- [53] B. Balle, G. Cherubin, and J. Hayes, “Reconstructing training data with informed adversaries,” 2022.
- [54] A.-V. Battis and L. Graner, “Risiken für die privatheit aufgrund von maschinellem lernen,” 2021.
- [55] F. Boenisch, “Privatsphäre und maschinelles lernen: Über gefahren und schutzmaßnahmen,” *Datenschutz und Datensicherheit - DuD*, vol. 45, pp. 448–452, 07 2021.
- [56] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, “Extracting training data from large language models,” *CoRR*, vol. abs/2012.07805, 2020.
- [57] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, “Stealing machine learning models via prediction apis,” 2016.
- [58] T. Orekondy, B. Schiele, and M. Fritz, “Knockoff nets: Stealing functionality of black-box models,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [59] L. Fan, K. W. Ng, C. Ju, T. Zhang, C. Liu, C. S. Chan, and Q. Yang, “Rethinking privacy preserving deep learning: How to evaluate and thwart privacy attacks,” *CoRR*, vol. abs/2006.11601, 2020.
- [60] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, “Membership inference attacks from first principles,” 2021.
- [61] C. Dwork, A. Smith, T. Steinke, and J. Ullman, “Exposed! a survey of attacks on private data,” *Annual Review of Statistics and Its Application (2017)*, 2017.
- [62] R. Joud, P. Moëllic, R. Bernhard, and J. Rigaud, “A review of confidentiality threats against embedded neural network models,” *CoRR*, vol. abs/2105.01401, 2021.
- [63] K. Sadeghi, A. Banerjee, and S. K. S. Gupta, “A system-driven taxonomy of attacks and defenses in adversarial machine learning,” *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, no. 4, pp. 450–467, 2020.

-
- [64] J. Stock, T. Petersen, C.-A. Behrendt, H. Federrath, and T. Kreutzburg, “Privatsphärefreundliches maschinelles lernen,” *Informatik Spektrum*, vol. 45, no. 2, pp. 70–79, 2022.
- [65] M. Veale, R. Binns, and L. Edwards, “Algorithms that remember: model inversion attacks and data protection law,” *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, p. 20180083, oct 2018.
- [66] C. Song and V. Shmatikov, “Overlearning reveals sensitive attributes,” 2020.
- [67] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel, “Extracting training data from large language models,” in *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, USENIX Association, Aug. 2021.
- [68] S. Gambs, A. Gmati, M. Hurfin, and D. Zekrifa, “Reconstruction attack through classifier analysis,” vol. 7371, pp. 274–281, 07 2012.

A Expanded Taxonomy

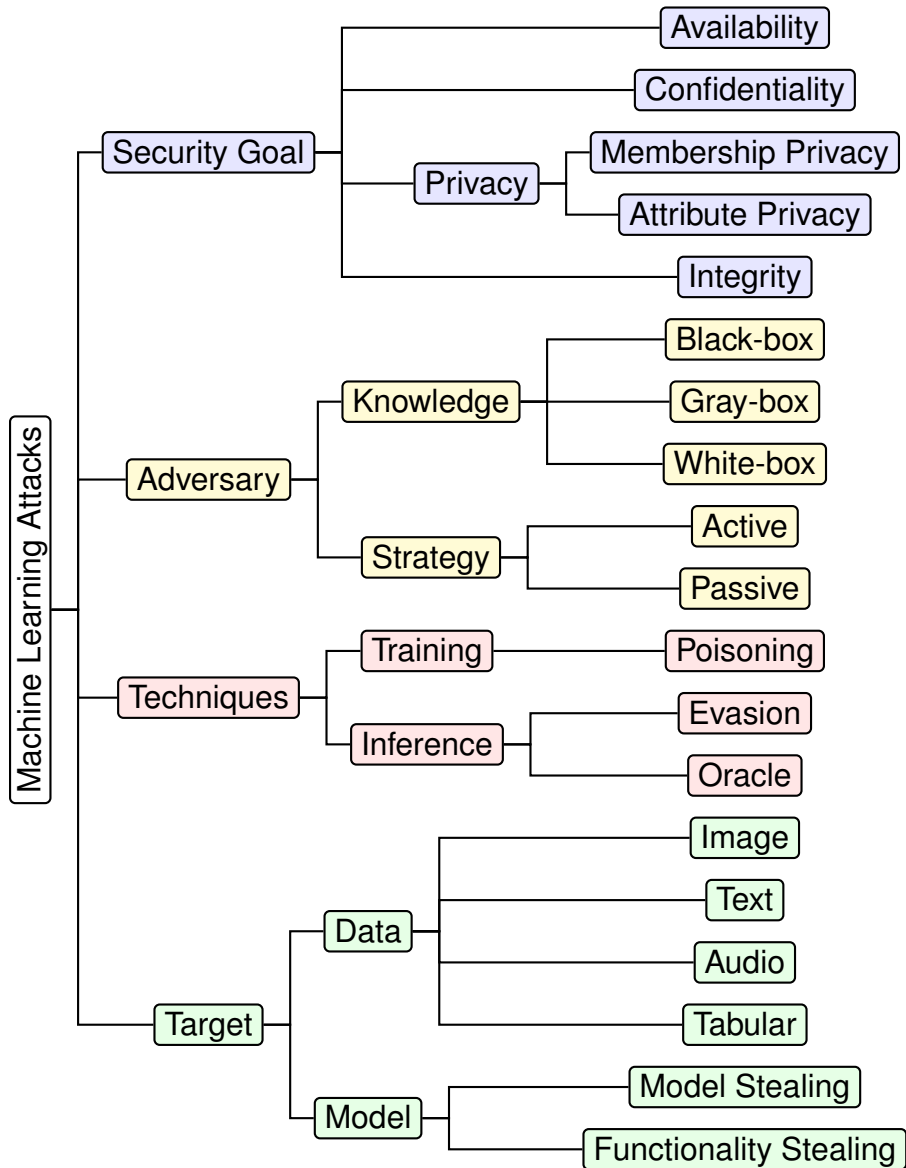


Figure 9: Example for possible expanded taxonomy.

B Expanded Classification

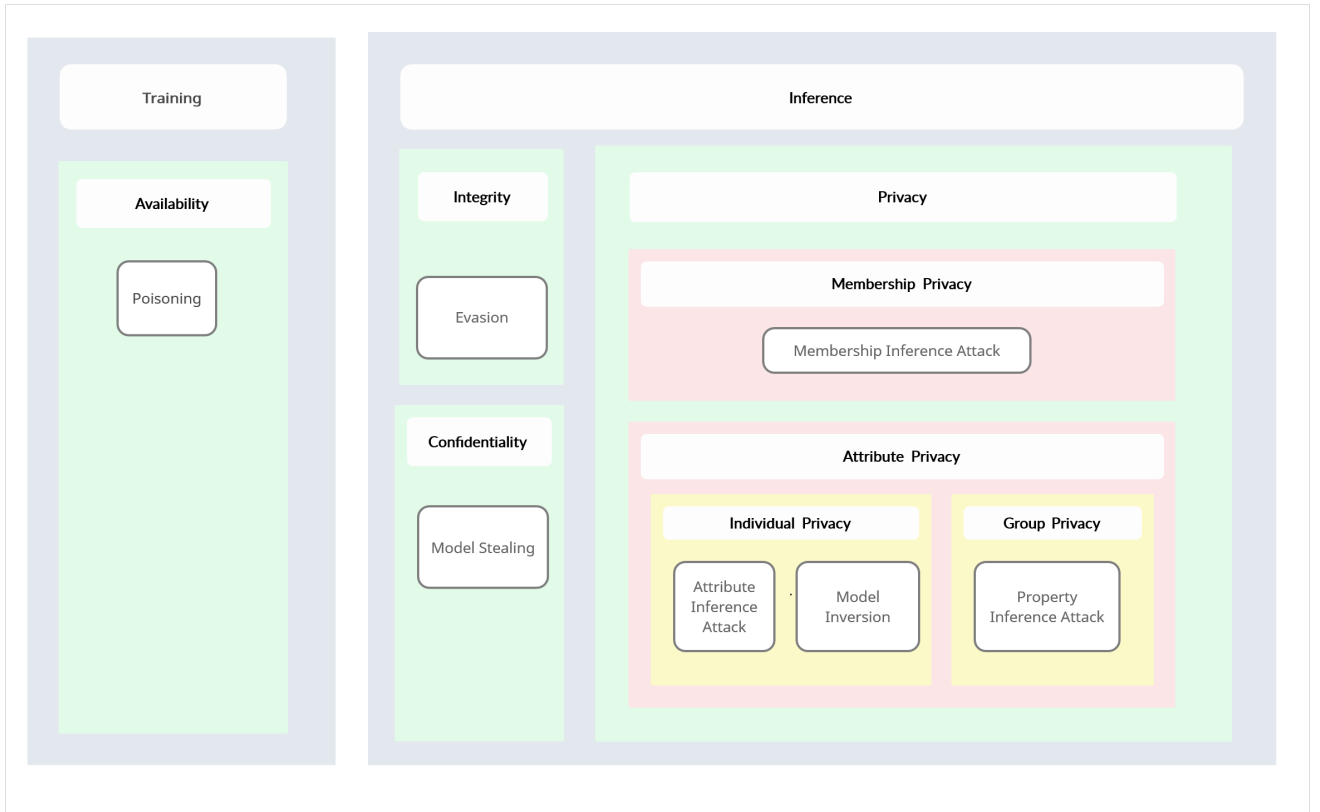


Figure 10: Classification of ML-attacks regarding timing and security goals