SoK: Model Inversion Attack Landscape: Taxonomy, Challenges, and Future Roadmap

Sayanton V. Dibbo*
* Department of Computer Science, Dartmouth College, Hanover, NH 03755, USA
* sayanton.v.dibbo.gr@dartmouth.edu

2023 IEEE 36th Computer Security Foundations Symposium (CSF) | 979-8-3503-2192-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSF57540.2023.00027

***Abstract*—A crucial module of the widely applied machine learning (ML) model is the model training phase, which involves large-scale training data, often including sensitive private data. ML models trained on these sensitive data suffer from significant privacy concerns since ML models can intentionally or unintendedly leak information about training data. Adversaries can exploit this information to perform privacy attacks, including *model extraction, membership inference,* and *model inversion*. While a model extraction attack steals and replicates a trained model functionality, and membership inference infers the data sample's inclusiveness to the training set, a model inversion attack has the goal of inferring the training data sample's sensitive attribute value or reconstructing the training sample (i.e., image/audio/text). Distinct and inconsistent characteristics of model inversion attack make this attack even more challenging and consequential, opening up *model inversion* attack as a more prominent and increasingly expanding research paradigm. Thereby, to flourish research in this relatively underexplored *model inversion* domain, we conduct the first-ever systematic literature review of the *model inversion* attack landscape. We characterize model inversion attacks and provide a comprehensive taxonomy based on different dimensions. We illustrate foundational perspectives emphasizing methodologies and key principles of the existing attacks and defense techniques. Finally, we discuss challenges and open issues in the existing model inversion attacks, focusing on the roadmap for future research directions.**

***Index Terms*—Model Inversion Attack, Taxonomy of Attack, Adversarial Capabilities, Defense, Sensitive Attribute**

I. Introduction

Artificial intelligence is the key driving force in our modern era, where machine learning (ML) and its applications play a significant role. In the current century, ML, particularly deep learning, is used almost everywhere, including predictive modeling [1], [2], social media analytics [3], [4], user authentication [5]–[8], image recognition [9], [10], audio analysis [11], [12], disease prediction and associated factors [13]–[17], and data analytics [18], [19]. However, these ML models can intentionally or unintendedly leak information about training data or memorize information during the model training [20], [21] phase, a crucial part of the ML model lifecycle. This phenomenon paves the way for different privacy attacks like model stealing [22], model inversion [23], and membership inference [24] attack. An adversary with different capabilities can perform these attacks with different aims, particularly targeting training data samples [25] which might contain sensitive private information like SSN, gender, racial identity, and facial expressions [26]–[28]. Consequently, these attacks can reveal sensitive information about an individual or generate fake images/audio, or even fool the model into providing incorrect predictions. Among these privacy attacks, model inversion is comparatively less explored [25], [29], particularly in the attribute inference (AI) sub-category. Model inversion attack (i.e., AI) has distinct and inconsistent characteristics compared to other privacy attacks, making it even more challenging [30].

ML models can leak training data information in different ways. For example, an adversary can query the target model to get a prediction reflecting the input-output data mapping. Also, an adversary can adapt techniques like an augmented synthetic dataset to query the model, which outputs prediction as memorized information from training data [31]. Eventually, the adversary can leverage those data to develop an attack model for estimating or reconstructing training samples [20]. Additionally, deep learning models can leak sensitive information in model weight updates or gradient parameters in federated learning environments [32]. Furthermore, models can leak sensitive training data during online learning [10], where the adversary can perform an attack leveraging the difference in output (prediction) before and after an update.

Protecting sensitive training data privacy has been a significant concern in the last few years [10], [23], [33]–[35]. Over the last few years, image and tabular data have been vulnerable to privacy attacks [10], [33], [35]. One direction of research explores possible privacy attacks against ML models [23], [35], [36] whereas another research direction investigates ways to defend against these attacks [23], [34]. Among different types of privacy attacks against ML models *Model Inversion (MI) attack* is the most challenging due to its inherent properties [30]. MI attacks and their defenses have been explored relatively recently [23], [34], which require a systematic review for better understanding and future pathways.

Model Inversion Attack and its Taxonomy: An adversary in MI attack intends to reconstruct training data or infer sensitive information in training data [33], [34]. For tabular training data, adversary infers sensitive attributes training samples [33], [37] or reconstructs images of individuals or generic images of a class in case of image data [10], [23]. We can classify MI attacks into two broader sub-classes: **1) Inference and 2) Reconstruction**. The first one refers to inferring sensitive attributes exactly or approximately or estimating properties related to the training data sample [33], [38]–[40]. **Inference** can be further categorized into three sub-categories depending

DOI 10.1109/CSF57540.2023.00027

on inference attack objective: **Attribute Inference (AI)** to infer exactly an individual's sensitive attribute values using output labels and other information like confidence scores and information about non-sensitive attributes (tabular data), depending on the attack strategy [23], [33], [36], whereas **Approximate Attribute Inference (AAI)** refers to inferring sensitive attribute values either approximately or close to the value of attributes in the training data sample [41], and **Property Inference (PI)** stands for inferring properties (not explicitly stated as an attribute in the training set) related to individual training data sample such as if an individual is wearing glasses or has a doctor's specialty [42], [43]. In the second category, i.e., **Reconstruction or Image Reconstruction (IR)**, an adversary applies image labels as inputs, along with some additional information like confidence scores, blurred or masked images, and objectively reconstructs images of an individual or a class representative [10]. However, an adversary can reconstruct specific individual images or a representative image of a class [35]. Therefore, *image reconstruction (IR)* can further be classified into two subtypes: Typical Image Reconstruction (TIR) [10], [23], [35], and Individual Image Reconstruction (IIR) [10], [23], [25], [35]. In Figure 1, we present a comprehensive taxonomy of different MI attacks.

Black-box vs. White-box Settings: There might be variations in terms of settings of the adversarial MI attack, i.e., the target model access might be a *black-box* or *white-box* [44] to an adversary. When the adversary has *black-box* access, it can get only query the target model but does not know details of the parameters like weights, gradients, etc., of the target model or any specific assumption underlying the model architectures. When it queries the model, it gets the output labels or confidence scores of classes [35], [45]. However, in *white-box* access, an adversary knows details about target model parameters and the entire transparent model architecture and can query the model or even capable of changing the model parameters [46], [47]. So, the adversary gets better control over the model in *white-box* setup and possibly performs MI and other privacy attacks with higher efficacy [35].

Our Contributions. We make the following contributions:

- We present a first-ever systematic literature review and characterization of the *model inversion* attacks and their defenses
- We provide a comprehensive taxonomy of the existing *model inversion* attacks, based on different criteria
- We illustrate foundational perspectives of the *model inversion* attacks in the literature, focusing on the key attack methodologies and underlying assumptions (in tables in the Appendix). We also outline state-of-the-art defense principles against model inversion attacks
- We highlight core challenges and discuss open issues in model inversion attack research domain

II. Preliminaries

ML models work on datasets to predict or make a decision based on a model trained by a dataset relevant to the problem domain. In the ML pipeline, first, the dataset is selected and preprocessed, and computed inputs (features) are used to train the ML model, which includes some sensitive and non-sensitive features as well in the training dataset. In the pipeline, after the model training step, a different validation dataset from the same distribution is used to validate the trained model and then used to test on test data to predict or make a decision on unknown data. MI attack architecture includes two trained/developed ML models (based on attack methodology). In the MI attack paradigm, the trained model is termed the *target* model, and the inversion model developed by the adversary is called the *attack* model [23]. In this section, we focus on types of ML models and learning techniques, the training process of ML models, MI attacks, and attack pipelines, as well as differential privacy– a commonly used defense technique for adversarial attacks against ML models.

A. ML models and Types of ML models

ML models are used to predict either discrete or continuous data. Commonly used ML models are classified into four types according to their technique and purposes, as follows:

Supervised Learning: This is a label-based ML model training approach. For example, if we have n samples of k features $x_1, x_2, ..., x_k$ in the training data, and there is a ground truth or label y_i for $i <= n$ associated with each of the n feature vectors. The ML model is trained on the input features corresponding to the output labels. So, the training step enables capturing the mapping between inputs (features) and outputs (labels) corresponding to each training sample [48]–[50]. When the training is complete, test data is applied, and for the input features, it predicts the class label for the feature vector. Different ML models like Support Vector Machines (SVM), Decision Trees, Naive Bayes, K-Nearest Neighbor (K-NN), Neural Networks, and Logistic Regression follow *supervised* technique [51], [52]. For discrete class labels, they are called as *classification* problem, and for continuous class labels, they are called *regression*.

Unsupervised Learning: This technique of ML uses unlabeled data to train the model. So, there are n training samples (X_{train}) of k features $x_1, x_2, ..., x_k$ in the training data, but no ground truth or label y_i. Therefore, in this learning, different *clustering* algorithms are adopted to align more similar training samples together. During testing, test samples X_{test} are checked with samples in the designed clusters to assign them to the cluster that is mostly similar. To measure the similarity between training (X_{train}) and test (X_{test}) samples, distance measures like l2 distance, manhattan distance, or cosine metrics are commonly used [49], [53]. These assigned clusters are considered as the particular class (predicted) to which the test samples belong. These techniques are used for pattern recognition, outlier detection, or anomaly detection. DBSCAN, K-means clustering are some of the ML algorithms of unsupervised learning [54].

Semi-supervised Learning: This technique is in between the above two methods, where some samples have labels associated with the feature space and some are unlabeled data.

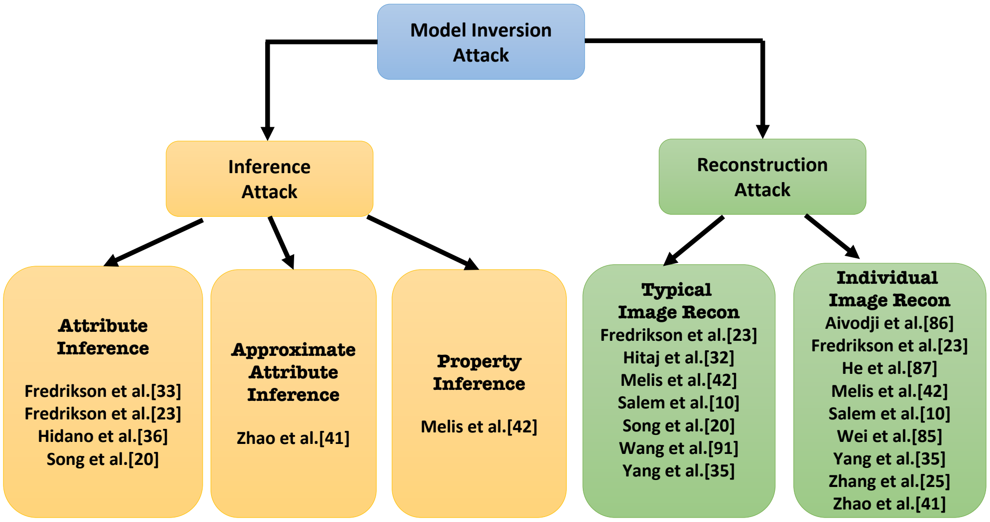


Fig. 1: A taxonomy of existing *Model Inversion (MI)* attacks in the literature, targeting different machine learning target models

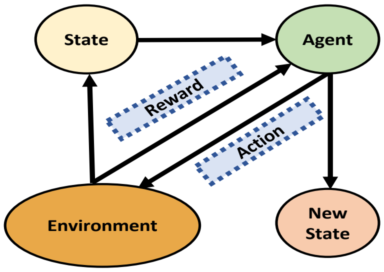


Fig. 2: Overview of Reinforcement Learning Technique

So the learning algorithm applies the knowledge about a few labeled samples (X_{label}) and tries predicting labels of the rest of the unlabeled dataset samples ($X_{unlabel}$) [55], [56]. A commonly used algorithm in this technique is self-learning, where a base classifier C_{base} is first trained with X_{label} and then queried with $X_{unlabel}$ samples to obtain pseudo labels $y_{unlabel}$, i.e., labels for unlabeled samples. Finally, a final model is trained with both labeled X_{label} and unlabeled data $X_{unlabel}$ with their pseudo labels $y_{unlabel}$. Learning concepts like few shot learning [57] and contrastive learning [58] follow semi-supervised based transfer learning fashion.

Reinforcement Learning: This is an iterative technique commonly used for Artificial Intelligence driven agent learning. In this technique, the *agent (A)* learns to take *action (a)* based on a given environment in such a way that maximizes the *reward (r)* [59]. So, there is *agent, reward, action* related to each *state (s)* of the agent interacting like a state transition diagram in *markov model*, as presented in Figure 2. In this technique, *states* are associated with *actions* taken by the *agent* in a way that maximizes the *reward*, i.e., minimal penalty. Finally, the set of *actions* $(a_1, ..., a_n)$ taken by *agent* construct the overall task. Another popular learning technology called *transfer learning* is booming in this era, where an ML model trained on large-scale datasets is used to test on a small dataset without further retraining. This training approach can also leak sensitive information on training data [60].

B. Model Inversion (MI) Attacks

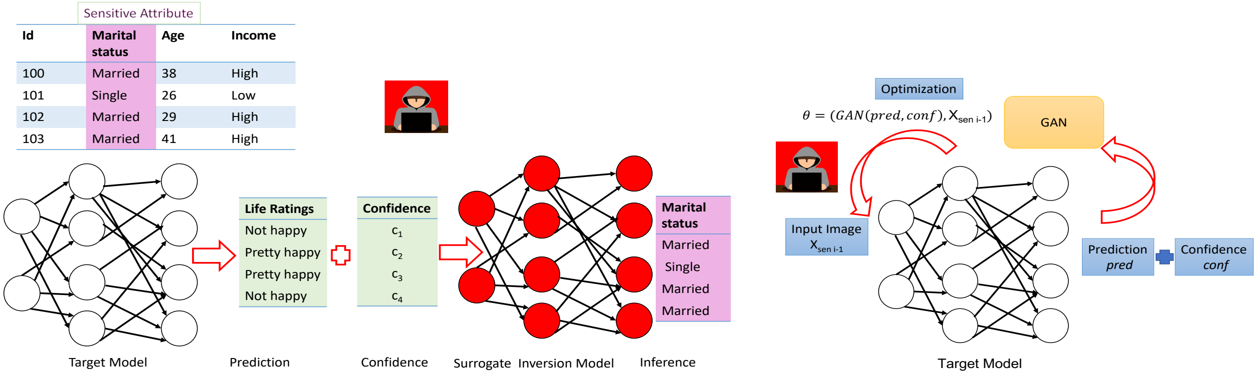
Model Inversion (MI) attacks leverage the output labels of the target model and additional auxiliary information such as confidence score, gradient or parameters of the model, etc., depending on implementation, for inverting the target model to infer training data sensitive attributes or reconstruct the sample of training data [29], [32], [33], [35]–[37], [61]. Suppose we have a dataset d containing m features and output labels y. We consider a classification problem with n possible classes in the output. Assume that, $x_1, x_2, ..., x_m$ are m input features in the dataset, where features $x_1, x_2, ..., x_t$ are sensitive features and features $x_{t+1}, ..., x_m$ are non-sensitive features. Therefore, the following function is the target model:

$$f_{tar} : R^m \rightarrow R^n \tag{1}$$

where R^m denote input features of m dimensions and R^n is the confidence values of n classes for each tuples of input features of d dimensions [23], [34], [36]. On the other side, the attack model will be the inverse of the target model, i.e.,

$$f_{atk} : [R^n, f_{tar}, y] \rightarrow R^s \tag{2}$$

where R^n denote the confidence score vector of n classes, f_{tar} is the target model, y is the predicted class in target model, and R^s denotes the sensitive attribute values in training dataset. We illustrate a detailed overview of the model inversion sensitive attribute inference (AI) attack in Figure 3a and image reconstruction (IR) attack in Figure 3b. An adversary can form the attack dataset D_{adv} performing row-wise column concatenation among prediction column y ('Life Ratings' in Figure 3a), inputs features $x_1, x_2, ..., x_m$ and highest class confidence score column $conf \in R^n$ ('Confidence' in Figure 3a), where all columns have the same number of rows. Then adversary can train the surrogate inversion model with D_{adv}



Id	Marital status	Age	Income
100	Married	38	High
101	Single	26	Low
102	Married	29	High
103	Married	41	High

(a) *Model Inversion Attribute Inference (AI) Attack*

(b) *Image Reconstruction (IR) Attack*

Fig. 3: (a.) Details of Model Inversion Attribute Inference Attack in *surrogate model training approach*. The target model takes the input dataset with 3 samples (on the left top), where *Marital status* is the sensitive attribute. The target model predicts *Life Ratings* either 'Not happy' or 'Pretty happy.' The adversary uses the predicted labels and confidence information (by row-wise column concatenation) to form D_{adv} (adversarial dataset) and train the surrogate inversion model to infer sensitive attributes, i.e., *Marital status* (shown as the output on the right side), (b.) Model Inversion attack for image reconstruction in *optimization-based approach*, where the adversary considers minimizing the optimization function θ iteratively (e.g., using GAN model with optimization algorithm) to obtain a better reconstruction of inputs/features ($X_{sen_{i-1}}$).

or perform an optimization-based technique to reconstruct inputs/features 3. We illustrate different inversion techniques in Section III-C.

C. Auxiliary Information

While performing the MI attack, the adversary might have a set of additional information, also called *auxiliary information* or *side-information*. This auxiliary information varies from model or implementation and depends on availability to the adversaries [36]. It serves as the attack model's additional input feature. Some of the *auxiliary information* include: output predicted labels y of the target classifier, confidence scores R^n of the n classes, partial or all non-sensitive attribute information training data, distributions of sensitive attributes, confusion matrix of target model C, and a dataset for the attacker to train [33], [35], [36]. In *TIR or IIR*, the *auxiliary information* may include a blurred image or masked image of the individual [10], [25], [35].

D. ML and Deep learning Model Training

Traditional ML (decision tree, SVM, KNN, etc.) [51], [52] and deep learning models like *convolutional neural network (CNN)* [62], [63], *auto encoder (AE)* [10], [64], and *generative adversarial network (GAN)* [65], [66] are trained to better learn the model parameters so that it can well predict when applied to unknown data samples. This process is called ML and deep learning model training. A few ML and deep learning model training techniques are discussed below:

- *Traditional ML Model Training:* The goal of training in supervised traditional ML models is to learn input-output dependency in training samples so that, when it gets the test samples, the prediction should be close to the actual expected value and *loss* function is minimized [67], [68]. The model with higher *loss* indicates bad training, and lower *loss* means a well-trained model. This loss function is usually measured by *squared error (SE)* [68], which can be expressed as:

$$SE = (y_{true} - y_{pred})^2 \tag{3}$$

where, y_{true} is the actual and y_{pred} is the predicted labels from the ML model. Another form of representation of this *loss* function is the *mean squared error (MSE)* [68], expressed as below:

$$MSE = \frac{1}{N} \sum_{(x,y) \in D_{test}} (y_{true} - y_{pred})^2 \tag{4}$$

where, N denotes total sample count, and D_{test} is the test dataset where each sample (x, y) belongs to this dataset.

- *CNN, GAN, and AE Training:* The CNN model is mainly used for image data, where the basic architecture include varying number of *convolution* and *pooling* layers [62], [69]. In the convolution layer, the image is filtered with the kernel to reduce the dimensionality, and padding is performed to maintain the same dimensions. After that, pooling is performed to get the maximal or minimal cells covering the stride positions like scanning. On the other hand GAN is composed of two different neural network: *discriminator (D)*, and *generator (G)* [32], [70], [71]. G is used to generate fake images, and the D matches the image produced by G with the original image and, based on that, provides the discrimination score back to G, as presented in Figure 5. In this iterative process, G regenerates images to match better, minimizing the discrimination score, and D becomes a fool with fake images to discriminate. AE is also a neural network based model, where the main components are the *encoder (E)* and *decoder (D)* [10], [64]. E performs mapping the

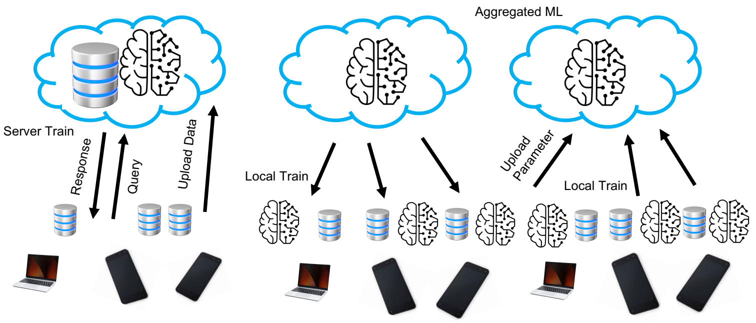


Fig. 4: Different ML modeling technique, from left to right: Centralized ML, where each client uploads data to server and server trains ML model, clients only can query; Distributed ML modeling, where clients train ML locally; Federated Learning modeling, where clients train ML on local data and uploads parameters to the server to aggregated ML model

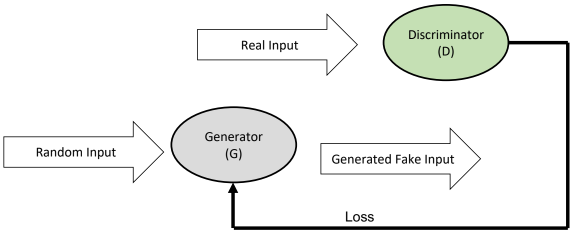


Fig. 5: Overview of the GAN Model

input to a latent space (i.e., hidden layer in a neural network), usually like a single layer perceptron; on the other hand, the D decodes the latent representation back to the original data (reconstructing images usually).
In all the various neural network-based models, the training goal is optimizing the *loss* function calculated as the difference between the prediction and expected outputs by iterative weight and bias adjustments [72]. There are different optimization methods that these models use, like adam, gradient descent, etc. [73], [74]. Although, among different optimization techniques, a particular gradient descent algorithm, named the *stochastic gradient descent (SGD)* is most commonly applied in applications. SGD follows the following updating formula:

$$\theta = \theta - \eta \nabla_{\theta} Cost(\theta; x, y) \tag{5}$$

where, θ is the updating parameter like weight, bias; ∇ is the parameter denoting gradient, η is learning rate, and $Cost(\theta; x, y)$ belongs to the iteratively optimized cost/loss function [74], [75].

- *Centralized, Distributed, and Federated Learning:* Depending on the server and client interaction involving training data, ML and deep learning training techniques can be categorized into the following three major categories: Centralized, Distributed, and Federated Learning [76]. In the centralized model, the ML model on central server connected with clients/devices is trained with the entire training dataset D_{train}. Clients only send data and may query for a specific service. In the distributed modeling, there are k clients and k distributed datasets, $D_1, ..., D_k$, where each client trains their own ML model using the corresponding local datasets [76], [77]. In the federated learning (FL) technique, the setup is different. Each client uses their local data to train their own models like distributed scheme; however, as opposed to the distributed one, in federated learning, clients keep data within the device and send only parameters (updates) to the central server periodically for aggregation tasks. This aggregated model might be shared with peers from the central server [32], [78]–[80]. FL can be further classified into two types: horizontal federated learning (HFL), where clients have same features space with different samples, and vertical federated learning (VFL), where sample space is identical but feature space is distinct among the clients [80]. The detailed process of these different learning approaches is illustrated in Figure 4.

E. Differential Privacy

Differential Privacy (DP) defines an algorithmic technique to represent the information about a dataset through a randomization procedure, which makes sure it prohibits disclosing any individual participant's identifiable information [33], [81]–[83]. This provides defense against membership inference attacks but is not very helpful against MI attack [34], since it compromises target model performances, i.e., there exists trade-offs between privacy and model performances. When privacy budget is higher and model performance is higher, the model is more susceptible to MI attacks.

III. Systematization of Model Inversion Attacks

Model inversion (MI) attack was first termed by Fredrikson et al. [33] into the privacy attack domain. The first methodological formalization of this attack in *white-box* and *black-box*– both scenarios are elaborated by Wu et al. [84]. Researchers have gradually explored more in MI attack techniques to infer sensitive attributes of training instances or reconstruct training instances leveraging auxiliary information. In this section, we systematize existing MI attacks, their objectives, foundational aspects, and key open issues with future research directions. We present a systematic summary of different MI attacks in Table I in the Appendix.

A. MI Attack Classification

As MI attacks aim to reconstruct training samples (e.g., image data) or estimate sensitive attribute information (e,g., tabular data), we can objectively classify MI attacks into two major categories: *(1) Inference*, and *(2) Reconstruction*, as discussed in Section I and illustrated in Figure 1. These MI attack techniques have a wide spectrum of variations and can be categorized depending on five primary characteristics or criteria. These categories are (i) based on target model access types, i.e., *white-box* or *black-box*, (ii) depending on inversion techniques incorporated, i.e., *training surrogate model* or *optimization-based approach*, (iii) the adversarial targeted types of data (image vs. tabular), (iv) learning schemes considered, i.e., *centralized* or *distributed*, or *federated* technique, and (v) availability of auxiliary information, i.e., *gradient information*, *confidence score information*, and *auxiliary dataset*. In Table I in the Appendix, we present the detailed characterization of the existing MI attacks.

B. MI Attack Objectives and Applications

An adversary in the model inversion attack aims to: (i) estimate sensitive attribute values in training instances or property of training instances, (ii) reconstruct training instances. Existing research approaches empirically investigate the first adversarial aim for tabular data [20], [23], [33], [36]. We categorize this attack scenario as the *attribute inference (AI)* attack (see Figure 1). This AI attack leads to the leaking of sensitive private data of an individual in the training set [33] and can have serious consequences, e.g., an adversary can target a patient disease prognosis model and can infer an individual's (training instances) previous medical records (attributes) or other sensitive information by posing this AI attack. In real life, machine learning-as-a-service providers (MLasS), e.g., Amazon, Google, and Microsoft, allow deploying ML models (trained on sensitive private data) in the clouds, which can be further queried with API provided by the MLasS. Adversaries with capabilities can leverage these APIs to perform AI attacks [22], [23], [33]. In recent research, researchers extend the MI attack to a relaxed version, i.e., *approximate attribute inference (AAI)* attack, which focuses on inferring the sensitive attribute value based on some acceptable threshold on the distance between the actual and inferred attribute values, i.e., the value that exists within the threshold [41]. Personalized medicine, lifestyle prediction, facial recognition, object identification, medical imaging, etc., are some of the applications of target models, where adversaries can target and perform MI attacks. In Table III in the Appendix, we present different MI attacks and their real-life applications.

C. Foundational Aspects of MI Attacks

MI attacks in the literature consider two basic inversion mechanisms as foundations– (i) *optimization-based approach*, and (ii) *surrogate model training approach*. In Table I in the Appendix, we present the approach considered in different MI attacks in the literature and their other characteristics.

- **Optimization-based approach:** In the *optimization-based approach*, the inversion purpose is turned into a gradient-based optimization problem that objectively reverses back the target model output to its input (reconstructed training samples), without training any additional model to handle this inversion task [23], [25], [35]. In this approach, the learning objective, i.e., optimization function in neural network-based models, is optimized in iterations in a way to generate better estimation or reconstruction of training instances. Existing research for reconstruction attacks that follow the *optimization-based approach* customizes the cost function in different ways to attain the goal for better reconstruction under different setups, e.g., white-box vs. black-box, the gradient vs. confidence score information availability, etc [25], [32], [33], [36], [85]. Fundamental steps in *optimization-based approach* are [23], [25]:

  ✓ Query the target model f_{tar} with auxiliary training set D_{aux} for the adversary (depending on setup/characteristics)

  ✓ Formulate the optimization function θ (similar to Eqn. 5) utilizing predictions $pred$, confidence scores $conf$ (if available from target model f_{tar}) and other adversarial capabilities (Figure 3b)

  ✓ Run an optimization algorithm for λ step size and α iterations; and in each iteration, compare the generated feature vector X_{sen_i} (reconstructed) to the feature representations in the last iteration $X_{sen_{i-1}}$ to update parameters according to function θ

  ✓ Perform post-processing, including rounding values and denoising (applying filters) reconstructed feature vectors X_{sen_i} to improve the quality of reconstruction. After α iterations, return the best-generated feature vector X_{sen_i}

  One of the commonly used optimization algorithms is the gradient descent with different variations like stochastic gradient descent (SGD), as illustrated in Section II-D. Irrespective of the algorithm, an important step in the *optimization-based approach* is to customize and formulate the optimization function function [23], [25], [36]. Existing research works consider different neural network models, including GAN [32], autoencoder, and multilayer perception (MLP) [23], with different optimization algorithms to inverse the target model back to the original inputs, i.e., reconstruct training samples.
- **Surrogate model training approach:** In the *surrogate model training approach*, an adversary exploits the basic auxiliary information and trains a surrogate model that leverages input-output correlation in the target model along with different auxiliary information and setups [20], [42], [86], [87] to inverse the target model (to estimate sensitive attributes or reconstruct training samples) by training surrogate model. An example of *surrogate model training approach* is illustrated in Figure 3a. This *surrogate model training approach* enables capturing the

input-output dependency in a better way compared to *optimization-based approach*, particularly for complex target models like long short-term memory (LSTM), convolutional neural networks (CNN), or recurrent neural network (RNN) [35]. The fundamental steps in the *surrogate model training approach* [10], [33], [35] are:

- ✓ Query the target model f_{tar} with auxiliary training set D_{aux} for the adversary (depending on setup/characteristics)
- ✓ Obtain predictions $pred$ and confidence scores $conf$ (if available) from target model f_{tar}
- ✓ Concatenate X_{in} with $pred$ and $conf$ (if available) to form the adversarial training dataset D_{adv}, where $X_{in} \in D_{aux}$ and X_{in} = X_{sen} + X_{nsen}, i.e., X_{in} is target model training samples' all attributes (both sensitive and non-sensitive attributes
- ✓ Train adversarial (i.e., inversion) model f_{adv} with D_{adv}, where inputs are all non-sensitive attributes X_{nsen}, predictions $pred$ and confidence scores $conf$ (if available); whereas output is the sensitive attribute value X_{sen}

In image reconstruction attacks, these sensitive attributes X_{sen} are features of reconstructed images from activation layers in neural networks [20], [32], [35]. In the literature, different MI attacks that consider *surrogate model training approach* follow these fundamental steps with some customization based on setup, auxiliary information availability, and for performance boostup [20], [35]. Note that the auxiliary training dataset D_{aux} might vary depending on available information [35]. For example, this set can be: i) the same as the training set of the target model, ii) a generic dataset related to a similar distribution (target model training data), or iii) a completely distinct dataset from any other source not necessarily identical to target model training data. Although performances might vary slightly, as expected, MI attacks are still effective with any type of auxiliary training set available [35].

D. Black-box MI Attacks

As discussed in Section I, *black-box* is a restricted access type to target models and hence the toughest setup for an adversary to perform a privacy attack like MI attacks [10]. An adversary with this access type does not have knowledge or control regarding the target model's internal architecture, parameters, weights, etc [32], [33], [36], [86]. The key insight for these MI attacks is to utilize the API access to the target model along with other capabilities to develop an inversion model through querying the target model for inferring/reconstructing sensitive attributes/training samples. We present a summary of the existing *black-box* and *white-box* MI attacks in Table III in the Appendix.

The first *black-box* MI attack was introduced by Fredrikson et al. [33] against a linear regression target model. This *black-box* attack only considered returned predictions from the target model to infer sensitive attributes. *Black-box* attack in [23] against the decision tree target model considered adversary can obtain both prediction and confidence scores. Different *black-box* attacks adopt various techniques to infer sensitive attributes depending on available information and setups. However, the basic steps involved in *Black-box* MI attacks are: (1) query the target model f_{tar} with data samples, (2) obtain predictions $pred$, confidence scores $conf$ based on setup, and (3) apply an algorithm to identify the best suitable candidate as the estimated sensitive attribute value.

Most commonly used algorithm in *step (3)* is the maximum a posterior (MAP) technique [23], [36]. Fredrikson et al. [23] in their MAP technique compute scores for each possible value of the sensitive attribute and return the one maximizing scores, where $score = c_{i,j} * p_i$ ($c_{i,j}$ is the value in the target model confusion matrix c_m, and p_i is class marginal prior) [23]. Similarly, Hidano et al. [36] compute the scores by multiplying class marginal p_i with an error term $e_i = err(y, \hat{x})$, where y is the actual value, and $\hat{x}$ is the sensitive attribute value considered. *Step (2)* assumptions also differ in existing *black-box* MI attacks. While most attacks assume only access to predictions $pred$ [33], [36], researchers also consider the availability of confidence scores $conf$ along with $pred$ and design the attacks utilizing all available information from target models [23].

Black-box MI attacks follow significantly distinct approaches in *Step (1)* depending on the setups. In [23], [33], the adversary is assumed to have knowledge of non-sensitive attributes of the training samples, thereby they query the target model simply by setting different values of sensitive attributes and follow *Step (3)* algorithm to obtain the best candidates. In another attack [36], the adversary injects poisoned samples by tweaking non-sensitive attribute values (making coefficients 0) to alter the target model f_{tar}. This controlled poisoning ensures minimum malicious samples and allows adversaries to have better control over the prediction $pred$. Finally, in *Step (3)*, it also applies a similar algorithm to find suitable candidates.

Adversaries might not have access to training samples (i.e., non-sensitive attribute values) [35]. Yang et al. [35] have implemented *black-box* MI attack (image reconstruction) without access to training samples. In *Step (1)*, this attack queries the target model with samples taken from generic distributions. In *Step (2)*, instead of having access to actual confidence scores $conf$, this work considers a more restricted scenario, where the adversary only gets truncated scores $conf_{trun}$ and applies this as inputs to design a surrogate inversion model (*Step (3)*) to obtain reconstructed images (input is $conf_{trun}$ and output is the features of reconstructed images). It also considers the other two sets of data in *Step (1)*, as described in Section III-C. This work shows the *black-box* MI attack is still successful even without full knowledge about target model training sets.

E. MI Attacks on Federated Learning

As data volume increases, expanding deep learning model computational power has become vital. Therefore, models are designed/deployed in *collaborative* fashion for both training [88] and inference [89], as illustrated in Figure 4. Among collaborative learning techniques, *federated learning (FL)* is the most promising because of its flexible and privacy-preserving multiparty updating principle (as discussed in Section II-D) [32], [42], [90].

While FL has been considered as privacy-preserving learning for a long, recent studies have shown it is also susceptible to privacy attacks like membership inference and model inversion attacks to some extent [32], [85], [87], [91]. In the FL setup, each client trains their local models with their private data and shares updates periodically to the server. However, this does not ensure the protection of private training data [32], [87], [91]. MI attacks against the FL clients focus on *image reconstruction (IR)* attacks and can be broadly classified into two major subcategories: (i) *malicious participant* and (ii) *malicious server*. In the *malicious participant* scenario, a malicious participant in FL acts as an adversary and tries to reconstruct training samples of other clients [32], [42], [85]. Whereas, in the *malicious server* scenario, the server itself acts as an adversary to reconstruct any participant's training samples [91].

Major steps in MI attacks in FL are: (a) target a specific clients' training data class/sample, (b) obtain gradient updates from the server (*malicious participant*), (c) utilize the gradient updates and other additional information to training an inversion model. Among the MI attacks in *malicious participant* subcategory, they differ in terms of objectives in *step (a)*, e.g., MI attacks in [32], [85], [90] targets a particular class (like class '4' in MNIST dataset [92], 'horse' in CIFAR-10 dataset [93], etc.) and reconstructs a typical image of that class, i.e., *typical image reconstruction (TIR)* attacks. Another type of MI attack has the goal of reconstructing an individual training image (sample) [42], [91] i.e., *individual image reconstruction (IIR)* attacks. *Step (b)* is applicable in *malicious participant* subcategory, where the malicious participant obtains the iterative gradient updates from the server (aggregator) and utilizes that to train an inversion model. For the *malicious server* scenario, the server already has access to the local updates [91]. In *Step (c)*, researchers commonly apply gradient-based techniques [32], [42], [85] to train the inversion models. Also, GAN models (discussed in Section II-D), GAN-based architectures/techniques [32], [91], or SGD-based models [42] can be trained utilizing gradient updates to reconstruct higher quality training data images.

An observation to note that all these different MI attacks in FL vary in terms of participant data setups [32], [42]. Most research considers data is disjoint among the clients, i.e., multiple clients have a distinct class of training data [32], [85], [90]. However, in a more realistic setup, it's common that multiple clients might share samples of the same class, i.e., not disjoint. This setup is considered in recent research, and it has been shown that even in this scenario, image reconstruction attacks are effective [42], [91].

F. MI Attacks in Online Learning

In general, training an ML model is considered expensive in terms of time and cost. Hence, with the availability of large-scale continuous data, retraining the model from scratch becomes a burden [10]. Therefore, *online training* has become an effective solution, which involves training an already trained ML model with only the updating dataset (new data samples). Suppose, $\mathcal{M}_{cur}$: $\mathcal{X}_{cur} \rightarrow \mathcal{Y}_{cur}$, where $\mathcal{X}_{cur}$ and $\mathcal{Y}_{cur}$ are the input and output sets that the current ML model is trained with. If $\mathcal{D}_{new}$ is the updating dataset (new data samples), then the *online training* process can be defined as $\mathcal{F}_{online}$: $\mathcal{M}_{cur} \rightarrow \mathcal{M}_{new}$, where $\mathcal{M}_{new}$ is the updated version of $\mathcal{M}_{cur}$ (trained with $\mathcal{D}_{new}$).

The *online training* process can also leak sensitive information on training samples or updating samples [10]. Salem et al. [10] designed MI attacks against the target model in *online training* setup to reconstruct samples in the update set $\mathcal{D}_{new}$. Fundamental steps in such an MI attack pipeline are: (i) select a $\mathcal{Q}_{prob}$ probing set and query the two versions of target models, i.e., $\mathcal{M}_{cur}$ and $\mathcal{M}_{new}$, (ii) utilize the posterior differences obtained from posterior probabilities in outputs of two target models in *step (i)*, (iii) train an inversion model to reconstruct training samples as outputs, taking posterior differences in *step (ii)* as inputs. The inversion model in *step (iii)* can be implemented using different neural network-based architecture, e.g., autoencoder used in [10], where the encoder (E) takes posterior difference as inputs and maps to intermediate vector representation, which is then decoded back to original samples by the decoder (D).

G. Memorization vs. MI Attacks

ML and deep learning models can be either benign or malicious [20]. A benign model does not *memorize* the training data during the model training phase. In contrast, an adversary can hide sensitive information (training dataset) in model training parameters. Adversaries can leverage this *memorized* information to pose privacy attacks, including membership inference and MI attacks [20], [21], [37], [46]. Therefore, memorization positively impacts risks for privacy attacks, i.e., the more a model can memorize, the more likely it is to be vulnerable to a privacy attack, e.g., an MI attack [37], [94]. One of the root causes for the memorization is the gap between model performance on training and test sets, measured by a popular term in ML called *overfitting* [95]. The more a model overfits, the more it loses generalizability and the more it memorizes, a way to leak training data sensitive private information [95]. Song et al. and Carlini et al., in their recent works, showed how the malicious models can *memorize* information regarding training data (image or text data) either intendedly [20] or unintendedly [21], which makes models vulnerable for adversarial attacks including membership inference [24] or MI attacks [33].

Unintended memorization refers to the inherent capability of an ML to hide training data dependency/correlation in model parameters, weights, biases, etc., during the iterations of the training phase [21]. An adversary, even with *black-box* access, can query these unintended memorized target models, and the memorization allows the models to leak sensitive information in the form of predictions $pred$ [21]. An adversary can leverage these predictions $pred$ to develop an inversion model and perform an MI attack, as discussed in Section III-C.

In ***intended memorization***, an adversary purposefully adopts techniques so that the target model is overparameterized during training to memorize training data [20]. One such technique is adding a regularization term with the usual loss function (e.g., SGD loss in Eqn. 5) for penalizing during optimization so that the model either (i) encodes sensitive attribute information to the signs of parameters or (ii) increase correlation between parameters and sensitive attributes [20]. Another way is to augment and increase the training set samples with the goal of encoding sensitive information directly to the least significant bits of parameters [20], which requires adversarial access to the model parameters to pose an MI attack against target models, i.e., *white-box* access.

H. Impact of Adversarial Knowledge on MI Attacks

Adversarial knowledge or capability plays a vital role in MI attack design and performances [23], [35]. In general, MI attacks designed using a similar inversion technique with more available adversarial capabilities achieve better attack performances/success rates compared to the scenario with less adversarial capabilities [36], [42], [86]. In Section II, we illustrate different auxiliary information and categorize auxiliary information considered in different MI attacks in Table I in the Appendix.

Among existing *black-box* attacks, MI attacks in [23], [35], [36] consider access to confidence scores $conf$ besides predictions $pred$ and is more effective than MI attack without access to $conf$ [33]. Other adversarial capabilities that Fredrikson et al. [23] consider are access to confusion matrix (C_m), class marginal priors, and all other non-sensitive attributes of training samples. Another *black-box* attack considered no or partial non-sensitive attribute information availability [35]. Although the adversary in *black-box* attack in [35] has partial adversarial knowledge on non-sensitive attributes, still this attack is effective compared to MI attack with full adversarial capabilities [23]. This is because *surrogate model training approach* inversion technique [35] is more robust and flexible compared to *optimization-based approach* (although involves overhead to train the surrogate model), while not compromising performance [35].

In the *white-box* setup, the impact of adversarial capabilities is even more prolific [25]. This is primarily because adversarial capabilities significantly contribute to gradient computations, which is the basic parameter that *white-box* setup makes available to an adversary. For example, *image reconstruction* attack designed in [25] considers different adversarial capabilities like access to blurred, T-masked, center-masked images or no auxiliary image. It shows that when no auxiliary image is available to an adversary, the reconstructed images are significantly different from the actual ground truth. The reconstruction quality drastically improves with the availability of any auxiliary information like blurred or masked images. Also, since center-masked images hide less sensitive parts in images compared to T-masked and thereby, reconstruction quality is comparatively higher in center-masked auxiliary information [25].

I. Open Issues and Future Directions

- **Attack with the minimal capabilities:** Existing MI attacks, both inference and reconstruction attacks consider a large pool of adversarial capabilities [23], [33], [35], [36], [86]. Some of these capabilities, e.g., access to confusion matrix, target model training samples' non-sensitive attribute values, or dataset to training inversion model, are not only overarching but also unrealistic to some extent. It is crucial to identify the minimal set of required capabilities for MI attacks and design such effective attacks under more realistic setups.
- **Performance stability in MI attacks:** Model inversion attacks aim to infer or reconstruct target models training samples. However, the same attack technique does not perform equally against all target models. Target model architecture might impact attack success rates. This opens up directions for further investigation on factors affecting performance stability in MI attacks across target models and to design more target model agnostic attacks, mitigating the stability factors.
- **Access type invariant attacks:** An avenue for future research is to introduce robust attacks that can be applied to either of the target model access types, i.e., *black-box* or *white-box*, without compromising attack performance significantly. State-of-the-art MI attacks are generally designed for particular target model access types [10], [33], [36]. While some attacks are customizable to suit other access types, they significantly suffer in performance [20], [23], [87].
- **Generalization vs. MI attack performances:** When an ML model is overparameterized during the training phase, it reduces generalizability as it tends to memorize more [21]. In practice, generalization is measured as a *gap* in model performance between training and test sets. If this *gap* is not significant, it implies the model generalizes well. Memorization and generalization are treated as two sides of the coin. A positive association is established between the MI attack and memorization [37], [94]. However, the empirical establishment of a relationship between generalization and MI attacks is yet to analyze.
- **Unified comparison metrics:** In current privacy and security research, there is no unified suitable metric for attack performance measures (different metrics are presented in Table III in the Appendix). For each primary category of MI attacks, a particular metric might capture the performance in a better way compared to other metrics. Therefore, based on category, in-depth investigation is a call for unfolding such a unified metric that all designed attacks should consider for experimental evaluations.
- **Reduced dependency on priors:** One interesting observation is that existing attacks are highly dependent on training data class marginal priors computed as a

ratio of samples that belong to a particular class and total samples. It is unexplored whether a model inversion attack is not effective without such information on target model training data, e.g., when an adversary generates its own synthetic data and performs such an attack.

- **Multimodal data-based MI attacks:** Model inversion attacks are focused on an image or tabular data mostly in the literature [23], [32], [35]. However, other data domains like text or audio/speech might be even more vulnerable and consequential due to privacy implications. One possible future direction is to study this attack in those domains as well. Also, some target models are trained with multimodal data leveraging different sources/data types. Existing attacks have not investigated the challenges of attacking target models trained with multimodal datasets.
- **Federated unlearning vs. MI attacks:** While researchers have investigated privacy leakage in FL and designed MI attacks against such paradigm [32], [42], [91], it has only been analyzed superficially and even unstudied. One such learning is vertical federated learning (VFL) [96], discussed in Section II-D. MI attacks are yet to design against this promising research direction in VFL. Also, in FL, after some iterations, any client might go down or remove, captured by a popular notion called *federated unlearning* [97]. It is still a question of whether an adversary can perform an MI attack, even in such scenarios.

IV. Defenses Against Model Inversion Attacks

Defense mechanisms against adversarial attacks have been comparatively less investigated in existing works. In this section, we illustrate the basic fundamentals of different defense techniques in the literature and describe open issues or challenges for the future. We present the foundational characteristics of different defense techniques against MI attacks in Table IV in the Appendix.

A. Defenses against back-box MI Attacks

Black-box MI attacks consider very restricted setup, and therefore designing defenses against these attacks are even more challenging [23], [33]. Existing research considers different approaches, including adding noise, rounding confidence scores, and differential privacy (DP) [23], [29], [33], [98].

1) Noise Superposition: Adding random noise to the posteriors, i.e., confidence scores, can work well if the adversary relies on confidence scores to design the attack model. A number of research approaches utilize this noise superposition technique for defending against MI attacks [10], [25], [98]. For example, suppose X_1 is a training sample and Y_1 is the posterior class probabilities, then adding $\tilde{\delta}$ noise to posterior class probabilities would result in updated posteriors Y_{up}= Y_1 + $\tilde{\delta}$. This defense imposes noises randomly from the uniform distribution to the posteriors when queried. This noise addition should be done on each sample individually so that the posteriors (on all queries) do not leak input-output correlation. Weak correlations between inputs and outputs reduce MI attack performances, as illustrated in empirical evaluations [10].

2) Perturbation and Rounding based Defenses: Among various defense techniques introduced against MI attacks, this one is investigated in [23], [29]. This technique involves perturbing or rounding the target model confidence scores. The defense can be *guided* or *unguided*. In the *unguided* perturbation or rounding, the target model randomly perturbs or simply rounds the confidence scores before making them available as outputs [23]. As a result, an adversary that implements the MI attack utilizing the confidence scores (obtained by querying the target model) would have incorrect predictions, i.e., sensitive attribute estimation or reconstructing samples. *Guided* perturbation or rounding serves a similar goal, only in a guided way, e.g., using a module that serves as the purifier of confidence score, which necessarily perturbs/clusters targeted confidence scores while minimizing loss function [29]. This clustering reduces dispersion and makes confidence scores indistinguishable (to samples), causing incorrect predictions for adversarial estimation.

3) Differential Privacy (DP) based Defenses: Differential privacy (DP) is a randomization technique considered to ensure training data privacy [33]. This technique commonly uses the ε parameter as the privacy budget, which indicates the maximum distance between sample values (ground truth) and their randomized values (DP randomized). Suppose X_{in} is the set of input samples (all attributes), f_{tar} is the target model trained with DP, so the randomized set of samples (differentially private set), X_{rnd}= $f_{tar}(X_{in})$ + $L(X_{in}, \varepsilon)$, where $L(X_{in}, \varepsilon)$ is the Laplacian distribution noise applied on samples with privacy budget ε. When ε is very small, varying inputs produce similar outputs; therefore less effective for an adversary to exploit privacy leakage. So, in general, DP is a great defense against privacy attacks. However, DP only randomizes dataset samples and prevents individual samples' inclusivity estimation in the dataset. However, it does not ensure attribute level privacy, which is the goal of MI attacks, and so DP is less effective against MI attacks [33], [34].

B. Minimizing Input-Output Dependency

One of the root causes for MI attacks is the dependency between inputs and outputs that the adversary leverages (through querying the target model) while designing attack [34]. Therefore, Wang et al. [34] proposed a new model-agnostic defense using *mutual information regularization* to reduce the input-output dependency. The key idea is to include an additional regularizer term using mutual information [99], [100] $\mathcal{I}(X_{in}, \hat{Y})$= $\mathcal{H}(\hat{Y})$ − $\mathcal{H}(\hat{Y}|X_{in})$ between inputs and outputs; along with traditional cross-entropy loss function $\mathcal{L}(y, f(X_{in}))$, where X_{in} and $\hat{Y}$ denote input features and output labels. This customized loss function with the additional regularizer penalty term enables iteratively reducing the *mutual information* through penalization and updating parameters iteratively [99], [100] between inputs and outputs.

C. Open Issues and Future Directions

- **Defending MI attacks in FL:** Federated learning has become a popular form of learning technique in recent years. Researchers have investigated privacy leakage in FL and designed MI attacks under different assumptions [32], [85], [91]. However, these attacks do not empirically analyze possible defenses to mitigate the attacks. Therefore, one big challenge is still remaining to design effective defenses against MI attacks in FL.
- **Target model agnostic defenses:** In general, model inversion attack defense is comparatively less explored [23], [34]. However, most of these defenses, e.g., rounding confidence in a neural network or priority-based sensitive attribute placement in decision tree [23], are targeted for specific target models. While some of these can be applied to other target models through customizing and tweaking [34], more generic forms of defenses agnostic to target models are yet to introduce and evaluate their efficacy.
- **Defense vs. target model utility:** There is a tradeoff between target model performance (utility) and defense success rates [33], [34]. To be more precise, as the defense techniques against MI attacks are applied, and they become successful in reducing attack performance, at the same time, the target models suffer much from downstream task performances, e.g., using DP as a defense [34]. It has become important to design effective MI attack defenses that do not compromise target model performances significantly.
- **Generalizable defense framework:** One of the challenges in privacy attack defenses is to implement a robust and generalizable defense framework that supports effective defenses against any privacy attack under any general assumptions. So far, in the literature, privacy attack defenses are ad-hoc, based on attack technique, downstream target model task (classification or regression), assumptions, target model access types, etc., which require further investigations to identify generic characteristics that bring them all privacy attacks under the same defense framework.
- **Adaptive Multifactor defense:** Studies found that input-output correlation in training data is a contributing factor for MI attacks [34]. There might be other factors that holistically control MI attack success. Also, different viable factors might impact differently based on setups like sample size, model parameter weight, data types, binary/multiclass classification tasks, etc. Therefore, understanding these factors and designing an adaptive defense based on the findings is a potential future direction.

V. Discussions and Future Work

In this section, we highlight takeaways in model inversion attack research domains, shortcomings in existing research, and potential ways to overcome challenges and improve attack/defense performances or future research directions.

- **Robust model inversion attacks:** Model inversion attack is still in flux. More investigations are required to design effective attacks under more realistic assumptions. This includes identifying minimal required capabilities for this attack under different setups and access types. It is also important to ensure these attacks are more robust against state-of-the-art defenses and invariant to target model architectures. Another research direction is to explore these attacks on scenarios where target models are trained in fairly recent learning concepts like zero short, few shot [57], and contrastive learning [58], besides federated learning techniques and different scenarios of FL like federated unlearning, as discussed in Section III-I.
- **Generalized defense against inversion attacks:** Considering its security and privacy threats, it has become a challenge to design effective and target model agnostic defenses against MI attacks. This requires further investigations in the federated learning environment, considering its large-scale applications. Also, in the future, as data volume and modality would increase, an important step is to identify factors controlling the MI attack and its relationship with other privacy attacks. This would enable the implementation of more generic multifactor-based privacy attack defense frameworks.
- **Multimodal MI attacks:** While data volume is increasing, data modality is also ever-growing. This phenomenon has made designing MI attacks on multimodal training data an open avenue for investigation. Since existing attacks are only limited to image and tabular data, they can further be explored in other domains like audio or text. Even it would be interesting to explore whether MI attacks are effective when target models are trained with multimodal data fusion techniques from multiple sources.

VI. Conclusions

Versatile AI and ML applications and large volume multi-modal data availability are the root causes for data privacy threats, either tabular/image/audio/text data domains. One such consequential threat is the model inversion attack, which objectively looks for inferring training data sensitive attributes (tabular data) or reconstructing training data samples of an individual/class (image/audio/text data). In recent years, researchers have introduced MI attacks exploiting different auxiliary information to infer sensitive attributes (tabular data), although most of them focus on reconstruction attacks (image data). This can further be extended to other modalities like against multi-modal audio/text data (both centralized and federated learning), even to rigorously explore inference attacks on tabular data. Additionally, effective generalized robust attack techniques are yet to investigate. Likewise, target model agnostic defenses against MI attacks are crucial. This paper provides a systematization of the MI attacks– a taxonomy of approaches, foundational aspects, open challenges, and potential future directions in the MI attack domain.

References

[1] J. B. Heaton, N. G. Polson, and J. H. Witte, "Deep learning for finance: deep portfolios," *Applied Stochastic Models in Business and Industry*, vol. 33, no. 1, pp. 3–12, 2017.

[2] S. Vhaduri, S. V. Dibbo, and C.-Y. Chen, "Predicting a user's demographic identity from leaked samples of health-tracking wearables and understanding associated risks," in *2022 IEEE 10th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2022, pp. 309–318.

[3] M. McGuirk, "Performing social media analytics with brandwatch for classrooms: a platform review," 2021.

[4] A. Subroto and A. Apriyana, "Cyber risk prediction through social media big data analytics and statistical machine learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–19, 2019.

[5] S. V. Dibbo, W. Cheung, and S. Vhaduri, "On-phone cnn model-based implicit authentication to secure iot wearables," in *The Fifth International Conference on Safety and Security with IoT*. Springer, 2023, pp. 19–34.

[6] S. Vhaduri, S. V. Dibbo, and W. Cheung, "Hiauth: a hierarchical implicit authentication system for iot wearables using multiple biometrics," *IEEE Access*, vol. 9, pp. 116 395–116 406, 2021.

[7] S. Vhaduri, W. Cheung, and S. V. Dibbo, "Bag of on-phone anns to secure iot objects using wearable and smartphone biometrics," *IEEE Transactions on Dependable and Secure Computing*, 2023.

[8] A. Muratyan, W. Cheung, S. V. Dibbo, and S. Vhaduri, "Opportunistic multi-modal user authentication for health-tracking iot wearables," in *The Fifth International Conference on Safety and Security with IoT: SaSeIoT 2021*. Springer, 2022, pp. 1–18.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[10] A. Salem, A. Bhattacharya, M. Backes, M. Fritz, and Y. Zhang, "Updates-leak: Data set inference and reconstruction attacks in online learning," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 1291–1308.

[11] P. Mahana and G. Singh, "Comparative analysis of machine learning algorithms for audio signals classification," *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 15, no. 6, p. 49, 2015.

[12] M. Y. Tachbelie, S. T. Abate, and T. Schultz, "Development of multilingual asr using globalphone for less-resourced languages: The case of ethiopian languages." in *INTERSPEECH*, 2020, pp. 1032–1036.

[13] S. Yang, J. M. S. Bornot, K. Wong-Lin, and G. Prasad, "M/eeg-based bio-markers to predict the mci and alzheimer's disease: a review from the ml perspective," *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2924–2935, 2019.

[14] S. V. Dibbo, Y. Kim, and S. Vhaduri, "Effect of noise on generic cough models," in *2021 IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2021, pp. 1–4.

[15] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[16] S. Morton, R. Li, S. Dibbo, and T. Prioleau, "Data-driven insights on behavioral factors that affect diabetes management," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2020, pp. 5557–5562.

[17] S. Vhaduri, S. V. Dibbo, and Y. Kim, "Environment knowledge-driven generic models to detect coughs from audio recordings," *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 4, pp. 55–66, 2023.

[18] W. Li, Y. Chai, F. Khan, S. R. U. Jan, S. Verma, V. G. Menon, X. Li *et al.*, "A comprehensive survey on machine learning-based big data analytics for iot-enabled smart healthcare system," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 234–252, 2021.

[19] A. M. S. Osman, "A novel big data analytics framework for smart cities," *Future Generation Computer Systems*, vol. 91, pp. 620–633, 2019.

[20] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, 2017, pp. 587–601.

[21] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *28th USENIX Security Symposium (USENIX Security 19)*, 2019, pp. 267–284.

[22] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis," in *25th USENIX security symposium (USENIX Security 16)*, 2016, pp. 601–618.

[23] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 1322–1333.

[24] N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang, "Membership privacy: a unifying framework for privacy definitions," in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 2013, pp. 889–900.

[25] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 253–261.

[26] X. Yang, Y.-Z. Wang, B. Wang, and G. Yu, "Privacy preserving approaches for multiple sensitive attributes in data publishing," *CHINESE JOURNAL OF COMPUTERS-CHINESE EDITION-*, vol. 31, no. 4, p. 574, 2008.

[27] V. S. Susan and T. Christopher, "Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes," *SpringerPlus*, vol. 5, no. 1, pp. 1–21, 2016.

[28] M. Wu, X. Zhang, J. Ding, H. Nguyen, R. Yu, M. Pan, and S. T. Wong, "Evaluation of inference attack models for deep learning on medical data," *arXiv preprint arXiv:2011.00177*, 2020.

[29] Z. Yang, B. Shao, B. Xuan, E.-C. Chang, and F. Zhang, "Defending model inversion and membership inference attacks via prediction purification," *arXiv preprint arXiv:2005.03915*, 2020.

[30] Y. Liu, R. Wen, X. He, A. Salem, Z. Zhang, M. Backes, E. De Cristofaro, M. Fritz, and Y. Zhang, "$ml - doctor$: Holistic risk assessment of inference attacks against machine learning models," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022, pp. 4525–4542.

[31] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[32] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the gan: information leakage from collaborative deep learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 603–618.

[33] 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*, 2014.

[34] T. Wang, Y. Zhang, and R. Jia, "Improving robustness to model inversion attacks via mutual information regularization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, 2021, pp. 11 666–11 673.

[35] Z. Yang, J. Zhang, E.-C. Chang, and Z. Liang, "Neural network inversion in adversarial setting via background knowledge alignment," in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019, pp. 225–240.

[36] 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)*. IEEE, 2017, pp. 115–11 509.

[37] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*. IEEE, 2018, pp. 268–282.

[38] N. Z. Gong and B. Liu, "You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors," in *25th USENIX Security Symposium (USENIX Security 16)*, 2016, pp. 979–995.

[39] ——, "Attribute inference attacks in online social networks," *ACM Transactions on Privacy and Security (TOPS)*, vol. 21, no. 1, pp. 1–30, 2018.

[40] B. Mei, Y. Xiao, R. Li, H. Li, X. Cheng, and Y. Sun, "Image and attribute based convolutional neural network inference attacks in social networks," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 869–879, 2018.

[41] 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 (EuroS&P)*. IEEE, 2021, pp. 232–251.

[42] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 691–706.

[43] 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 *ACM SIGSAC conference on computer and communications security*, 2018, pp. 619–633.

[44] L. Pengcheng, J. Yi, and L. Zhang, "Query-efficient black-box attack by active learning," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 1200–1205.

[45] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[46] K. Leino and M. Fredrikson, "Stolen memories: Leveraging model memorization for calibrated white-box membership inference," in *29th USENIX Security Symposium*, 2020, pp. 1605–1622.

[47] 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, 2019, pp. 739–753.

[48] R. Krueger, J. Beyer, W.-D. Jang, N. W. Kim, A. Sokolov, P. K. Sorger, and H. Pfister, "Facetto: Combining unsupervised and supervised learning for hierarchical phenotype analysis in multi-channel image data," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 227–237, 2019.

[49] A. N. Khan, M. Y. Fan, A. Malik, and R. A. Memon, "Learning from privacy preserved encrypted data on cloud through supervised and unsupervised machine learning," in *2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. IEEE, 2019, pp. 1–5.

[50] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1–16, 2019.

[51] L. Sun, S. Fu, and F. Wang, "Decision tree svm model with fisher feature selection for speech emotion recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, pp. 1–14, 2019.

[52] J. Hu and J. Min, "Automated detection of driver fatigue based on eeg signals using gradient boosting decision tree model," *Cognitive neurodynamics*, vol. 12, no. 4, pp. 431–440, 2018.

[53] M. Chen, X. Shi, Y. Zhang, D. Wu, and M. Guizani, "Deep features learning for medical image analysis with convolutional autoencoder neural network," *IEEE Transactions on Big Data*, 2017.

[54] O. Limwattanapibool and S. Arch-int, "Determination of the appropriate parameters for k-means clustering using selection of region clusters based on density dbscan (srcd-dbscan)," *Expert Systems*, vol. 34, no. 3, p. e12204, 2017.

[55] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[56] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.

[57] S. Rahman, S. Khan, and F. Porikli, "A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5652–5667, 2018.

[58] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 8765–8775, 2020.

[59] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, J. Pineau *et al.*, "An introduction to deep reinforcement learning," *Foundations and Trends® in Machine Learning*, vol. 11, no. 3-4, pp. 219–354, 2018.

[60] S. P. Liew and T. Takahashi, "Faceleaks: Inference attacks against transfer learning models via black-box queries," *arXiv preprint arXiv:2010.14023*, 2020.

[61] X. Zhao, W. Zhang, X. Xiao, and B. Lim, "Exploiting explanations for model inversion attacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 682–692.

[62] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task cnn model for attribute prediction," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1949–1959, 2015.

[63] O. Ajao, D. Bhowmik, and S. Zargari, "Fake news identification on twitter with hybrid cnn and rnn models," in *Proceedings of the 9th international conference on social media and society*, 2018, pp. 226–230.

[64] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, "Grammar variational autoencoder," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1945–1954.

[65] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[66] M. Kahng, N. Thorat, D. H. Chau, F. B. Viégas, and M. Wattenberg, "Gan lab: Understanding complex deep generative models using interactive visual experimentation," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 310–320, 2018.

[67] B. Silva, F. R. Barbosa-Anda, and J. Batista, "Multi-view fine-grained vehicle classification with multi-loss learning," in *2021 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE, 2021, pp. 209–214.

[68] A. K. Fletcher, S. Rangan, and P. Schniter, "Inference in deep networks in high dimensions," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 1884–1888.

[69] K. Xia, J. Huang, and H. Wang, "Lstm-cnn architecture for human activity recognition," *IEEE Access*, vol. 8, pp. 56 855–56 866, 2020.

[70] P. F. de Araujo-Filho, G. Kaddoum, D. R. Campelo, A. G. Santos, D. Macêdo, and C. Zanchettin, "Intrusion detection for cyber–physical systems using generative adversarial networks in fog environment," *IEEE Internet of Things Journal*, vol. 8, no. 8, pp. 6247–6256, 2020.

[71] Y. Jo and J. Park, "Sc-fegan: Face editing generative adversarial network with user's sketch and color," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1745–1753.

[72] J.-Y. Kim and S.-B. Cho, "Evolutionary optimization of hyperparameters in deep learning models," in *2019 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2019, pp. 831–837.

[73] D. Marin, M. Tang, I. B. Ayed, and Y. Boykov, "Beyond gradient descent for regularized segmentation losses," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 187–10 196.

[74] H. Yu, R. Jin, and S. Yang, "On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization," in *International Conference on Machine Learning*. PMLR, 2019, pp. 7184–7193.

[75] L. Nguyen, P. H. Nguyen, M. Dijk, P. Richtárik, K. Scheinberg, and M. Takác, "Sgd and hogwild! convergence without the bounded gradients assumption," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3750–3758.

[76] S. A. Rahman, H. Tout, C. Talhi, and A. Mourad, "Internet of things intrusion detection: Centralized, on-device, or federated learning?" *IEEE Network*, vol. 34, no. 6, pp. 310–317, 2020.

[77] A. Jochems, T. M. Deist, J. Van Soest, M. Eble, P. Bulens, P. Coucke, W. Dries, P. Lambin, and A. Dekker, "Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept," *Radiotherapy and Oncology*, vol. 121, no. 3, pp. 459–467, 2016.

[78] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[79] P. Vanhaesebrouck, A. Bellet, and M. Tommasi, "Decentralized collaborative learning of personalized models over networks," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 509–517.

[80] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[81] C. Dwork, "Differential privacy: A survey of results," in *International conference on theory and applications of models of computation*. Springer, 2008, pp. 1–19.

[82] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy." *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211–407, 2014.

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

[84] 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)*. IEEE, 2016, pp. 355–370.

[85] W. Wei, L. Liu, M. Loper, K.-H. Chow, M. E. Gursoy, S. Truex, and Y. Wu, "A framework for evaluating gradient leakage attacks in federated learning," *arXiv preprint arXiv:2004.10397*, 2020.

[86] U. Aïvodji, S. Gambs, and T. Ther, "Gamin: An adversarial approach to black-box model inversion," *arXiv preprint arXiv:1909.11835*, 2019.

[87] Z. He, T. Zhang, and R. B. Lee, "Model inversion attacks against collaborative inference," in *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019, pp. 148–162.

[88] T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman, "Project adam: Building an efficient and scalable deep learning training system," in *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, 2014, pp. 571–582.

[89] S. Teerapittayanon, B. McDanel, and H.-T. Kung, "Distributed deep neural networks over the cloud, the edge and end devices," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2017, pp. 328–339.

[90] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," *International Journal of Security and Networks*, vol. 10, no. 3, pp. 137–150, 2015.

[91] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang, and H. Qi, "Beyond inferring class representatives: User-level privacy leakage from federated learning," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 2512–2520.

[92] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[93] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," *online: http://www. cs. toronto. edu/kriz/cifar. html*, vol. 55, p. 5, 2014.

[94] B. Wu, S. Zhao, G. Sun, X. Zhang, Z. Su, C. Zeng, and Z. Liu, "P3sgd: Patient privacy preserving sgd for regularizing deep cnns in pathological image classification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2099–2108.

[95] V. Feldman and C. Zhang, "What neural networks memorize and why: Discovering the long tail via influence estimation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2881–2891, 2020.

[96] S. Feng and H. Yu, "Multi-participant multi-class vertical federated learning," *arXiv preprint arXiv:2001.11154*, 2020.

[97] A. Halimi, S. Kadhe, A. Rawat, and N. Baracaldo, "Federated unlearning: How to efficiently erase a client in fl?" *arXiv preprint arXiv:2207.05521*, 2022.

[98] T. Titcombe, A. J. Hall, P. Papadopoulos, and D. Romanini, "Practical defences against model inversion attacks for split neural networks," *arXiv preprint arXiv:2104.05743*, 2021.

[99] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[100] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *International Conference on Machine Learning*. PMLR, 2018, pp. 531–540.

[101] I. W. P. Consortium, "Estimation of the warfarin dose with clinical and pharmacogenetic data," *New England Journal of Medicine*, vol. 360, no. 8, pp. 753–764, 2009.

[102] W. Hickey, "FiveThirtyEight: How Americans Like Their Steak," http://fivethirtyeight.com/datalab/how-americans-like-their-steak/, 2014.

[103] J. Prince, "Social science research on pornography," http://byuresearch.org/ssrp/downloads/GSShappiness.pdf.

[104] G. Research, "MovieLens 1M Dataset," http://grouplens.org/datasets/movielens/, 2003.

[105] F. S. Samaria and A. C. Harter, "Parameterisation of a stochastic model for human face identification," in *Proceedings of 1994 IEEE workshop on applications of computer vision*. IEEE, 1994, pp. 138–142.

[106] H.-W. Ng and S. Winkler, "A data-driven approach to cleaning large face datasets," in *2014 IEEE international conference on image processing (ICIP)*. IEEE, 2014, pp. 343–347.

[107] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[108] K. Lang, "Newsweeder: Learning to filter netnews," in *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 331–339.

[109] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, 2011, pp. 142–150.

[110] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," *Advances in neural information processing systems*, vol. 29, pp. 2234–2242, 2016.

[111] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3730–3738.

[112] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[113] N. Zhang, M. Paluri, Y. Taigman, R. Fergus, and L. Bourdev, "Beyond frontal faces: Improving person recognition using multiple cues," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4804–4813.

[114] B. Verhoeven and W. Daelemans, "Clips stylometry investigation (csi) corpus: A dutch corpus for the detection of age, gender, personality, sentiment and deception in text." in *LREC*, 2014, pp. 3081–3085.

[115] D. Yang, D. Zhang, L. Chen, and B. Qu, "Nationtelescope: Monitoring and visualizing large-scale collective behavior in lbsns," *Journal of Network and Computer Applications*, vol. 55, pp. 170–180, 2015.

[116] D. Yang, D. Zhang, and B. Qu, "Participatory cultural mapping based on collective behavior data in location-based social networks," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 7, no. 3, pp. 1–23, 2016.

[117] Y. O. D. Repository, "Yelp Dataset," https://www.yelp.com/dataset, 2004, [Online; accessed 12-July-2022].

[118] G. Chen, J. Zhou, and Z. Liu, "Global synchronization of coupled delayed neural networks and applications to chaotic cnn models," *International Journal of Bifurcation and Chaos*, vol. 14, no. 07, pp. 2229–2240, 2004.

[119] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

[120] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[121] M. Backes, M. Humbert, J. Pang, and Y. Zhang, "walk2friends: Inferring social links from mobility profiles," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1943–1957.

[122] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[123] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "Auc: a misleading measure of the performance of predictive distribution models," *Global ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, 2008.

APPENDIX

Paper selection Methodology: In this manuscript, we systematically study existing model inversion attacks since it was first introduced [33] till date. We choose papers as follows: *step1:* We consider paper [33] as the baseline. *step2:* We have done brute force searches in both defense and attack directions for the most influential works. *step3:* We expand the search radius in five dimensions (Table I in Appendix): (i) data types (image vs. tabular), (ii) target model access types (*black-box* vs. *while-box*), (iii) inversion technique (training vs. optimization) types, (iv) model learning (centralized, distributed, federated) types, and (v) auxiliary information (confidence-based, gradient-based, auxiliary data-based) types.

TABLE I: A Summary of the Systematization of Model Inversion (MI) Attacks against Target ML Models (*** *Infer=Inference, Recons=Reconstruction, Optim=Optimization-based Approach, Central=Centralized, Feder=Federated, Distri=Distributed, Confi=Confidence Score*)

Paper	Objective Type		Access Type		Inversion Technique		ML Modeling			Auxiliary Information		
	Infer	Recons	Black-box	White-box	Training	Optim	Central	Feder	Distri	Confi	Gradient	Data
Fredrikson et al. [33]	✓		✓			✓	✓					
Fredrikson et al. [23]	✓	✓	✓	✓		✓	✓			✓		
Hidano et al. [36]	✓		✓			✓	✓			✓		
Hitaj et al. [32]		✓		✓	✓	✓		✓			✓	
Song et al. [20]	✓	✓	✓	✓	✓	✓	✓				✓	
Aivodji et al. [86]		✓	✓		✓	✓	✓				✓	
Melis et al. [42]	✓		✓		✓	✓		✓			✓	
Wang et al. [91]		✓		✓	✓	✓		✓			✓	
Yang et al. [35]		✓	✓		✓		✓			✓		
He et al. [87]		✓	✓	✓	✓	✓			✓		✓	
Wei et al. [85]		✓		✓		✓		✓			✓	
Zhang et al. [25]		✓		✓		✓	✓					✓
Salem et al. [10]		✓	✓		✓		✓					✓
Zhao et al. [41]	✓		✓	✓	✓		✓			✓		

TABLE II: A Summary of Existing Model Inversion Attacks and their Properties

Paper	Attack Class	Attack Subcategory	Dataset	Performance Measure	ML Task	ML Model	Access Type	Application
Fredrikson et al. [33]	AI	Individual	IWPC [101]	Accuracy, AUCROC	Regression	Linear Regression	Black-box	Personalized Medicine
Fredrikson et al. [23]	IR and AI	Class Inference and Individual	FiveThirtyEight [102] and GSS [103]	Accuracy, precision, recall, % correct	Classification	Decision tree, Deep Neural Network	White-box, Black-box	Life-style choice, and Facial Recognition
Hidano et al. [36]	AI	Individual	FiveThirtyEight [102], and MovieLens [104]	# of Posining Samples, RMSE (target), Success Rates (Attack)	Classification	Linear Regression	Black-box	Product Recommendation, Lifestyle Prediction
Hitaj et al. [32]	IR	Class Inference	MNIST [92], and AT&T dataset of faces [105]	Accuracy	Classification	CNN	White-box	Image Reconstruction, Facial Recognition
Song et al. [20]	IR and AI	Class Inference	FaceScrub [106], CIFAR10 [93], LFW [107], 20 newsgroup [108], and IMDB [109]	Mean Abs Pixel Error (MAPE), Precision, Recall, Similarity	Classification	CNN, RES, SVM, LR	Black-box, White-box	Object Identification, Sentiment Analysis
Wang et al. [91]	IR	Class Inference	MNIST [92], and AT&T dataset of faces [105]	Inception Score [110]	Classification	CNN	White-box	Image Reconstruction, Object Identification
Yang et al. [35]	IR	Individual and Class Inference	FaceScrub [106], CelebA [111], CIFAR10 [93], and MNIST [92]	Accuracy, Avg. Reconstruction Loss	Classification	Deep Neural Network (CNN)	Black-box	Facial Recognition, Medical Imaging
He et al. [87]	IR	Individual	MNIST [92], and CIFAR10 [93]	PSNR, SSIM	Classification	Deep Neural Network (CNN)	White-box, Black-box	Object Identification

TABLE III: A Summary of Existing Model Inversion Attacks and their Properties (Continued)

Paper	Attack Class	Attack Subcategory	Dataset	Performance Measure	ML Task	ML Model	Access Type	Application
Aivodji et al. [86]	IR	Individual	MNIST [92], and Pilot parliament [112]	Fidelity, Acc, Global Score, Categorical Acc	Classification	Deep Neural Network (CNN) and Multilayer Perceptron (MLP)	Black-box	Object Identification, Facial Recognition
Melis et al. [42]	PI and IR	Individual and Class Inference	FaceScrub [106], LFW [107], PIPA [113], CSI [114], FourSquare [115], [116], Yelp-health [117], and Yelp-author [117]	AUC and Precision	Classification	Deep Neural Network (CNN [118], and RNN [119] model)	Black-box	Text data property inference, Facial Recognition
Wei et al. [85]	IR	Individual	MNIST [92], CIFAR10 [93], LFW [107], and CIFAR100 [93]	Attack Success Rate, MSE, SSIM, Attack Iteration	Classification	Deep Neural Network (CNN)	White-box	Person Identification, Object Detection
Zhang et al. [25]	IR	Individual	MNIST [92], ChestX-ray8 [120], and CelebA [111]	PSNR, Accuracy, Feat Dist, and KNN Dist	Classification	Deep Neural Network (CNN)	White-box	Person Identification, Medical Imaging
Salem et al. [10]	IR	Individual and Class Inference	MNIST [92], CIFAR10 [93], Insta-NY [121]	Accuracy, MSE, KL-divergence	Classification	Deep Neural Network (CNN)	Black-box	Person Identification, Image Identification
Zhao et al. [41]	AAI	Class Inference	FourSquare [115], [116], CIFAR10 [93], and Purchase100 [122]	Area Under Curve (AUC) score [123]	Classification	Logistic Regression, Support Vector Machines, Random Forests, and Neural Network	Black-box and White-box	Person Identification, Person/Object Property Inference

TABLE IV: A Summary of Different Defenses Against MI Attacks

Paper	Attack Class	Attack Subcategory	Dataset	Attack Performance Measure	ML Task	ML Model	Access Type	Defense Technique	Application
Fredrikson et al. [33]	AI	Individual	IWPC [101]	Inversion Accuracy	Regression	Linear Regression	Black-box	DP	Personalized Medicine
Fredrikson et al. [23]	IR and AI	Class Inference and Individual	FiveThirtyEight [102] and GSS [103]	Inversion Accuracy, % correct	Classification	Decision tree, Deep Neural Network	White-box, Black-box	Reducing Confidence Precision, Sensitive Feature Prioritization	Life-style choice, Medical diagnosis, and Facial Recognition
Yang et al. [29]	IR	Individual	FaceScrub [106], CIFAR10 [93], Purchase [122]	Classifier Accuracy, Inversion Error, Inference Accuracy, Confidence Score Distortion, and Training Time	Classification	Deep Neural Network	Black-box	Confidence Score Purification	Person Idetification, Facial Recognition
Wang et al. [34]	IR and AI	Individual	FaceScrub [106], CelebA [111], CIFAR10 [93], IWPC [101], FiveThirtyEight [102]	Accuracy, F-1, AUROC, L2 Distance, MSE	Classification, Regression	Deep Neural Network, Decision Tree, Linear Regression	White-box, Black-box	Mutual Information Regularization	Person Idetification, Medical Imaging, Life-style choice, Facial Recognition
Tom et al. [98]	IR	Individual	MNIST [92]	Accuracy	Classification	Deep Neural Network	Black-box	Laplacian Noise Defense	Object Identification