Membership Inference Attacks Against Machine Learning Models

ALUCH Yasmine
ENSAE Paris
yasmine.aluch@ensae.fr

OUALY Ossama
ENSAE Paris
ossama.oualy@ensae.fr

SAIMANE Nawal
ENSAE Paris
nawal.saimane@ensae.fr

Abstract

Machine learning models, particularly deep neural networks, are vulnerable to membership inference attacks (MIAs) that can reveal whether specific data points were used during training, posing significant privacy risks in sensitive applications. This study systematically evaluates three attack strategies—confidence-based, loss-based, and shadow model-based—against convolutional neural networks trained on CIFAR-10. We implement a realistic gray-box scenario using a convolutional autoencoder to generate synthetic training data, simulating an attacker with limited knowledge of the target distribution. Our experiments show that threshold-based attacks achieve AUC 0.67, while shadow models trained on synthetic data reach AUC 0.62, both significantly above the 0.50 random baseline. The strong correlation between overfitting (26.4% generalization gap) and attack success confirms that memorization increases vulnerability, underscoring the need for privacy-preserving techniques in practical deployments.

1 Introduction

Deep neural networks have transformed critical applications in healthcare, finance, and security, yet their tendency to memorize training data creates serious privacy vulnerabilities. Membership inference attacks (MIAs) exploit this memorization: an attacker querying a deployed model via an API can infer whether specific data points were used during training. Such attacks can expose sensitive information—medical models may reveal patient records, facial recognition systems may disclose individuals' identities—making privacy protection essential in computer vision and other domains.

This study evaluates three MIA strategies against CIFAR-10 image classifiers. **Confidence-based attacks** exploit higher prediction certainty on training samples. **Loss-based attacks** leverage lower cross-entropy loss on memorized data. **Shadow model attacks** train auxiliary models to learn membership patterns from output distributions. Crucially, we simulate realistic constraints where attackers lack access to real training data, instead generating synthetic images via autoencoders.

Our key findings: (1) simple threshold methods achieve AUC 0.67 (accuracy 63%), above the 50% random baseline; (2) shadow models with synthetic data reach AUC 0.64 (accuracy 57.9%); (3) a 26.4% generalization gap strongly predicts attack success, confirming that overfitting drives privacy leakage. These results demonstrate that deploying overfitted models on sensitive data requires robust privacy safeguards.

Code availability: Our implementation is publicly available on GitHub Saimane et al. [2026].

2 Related Work

Membership inference attacks (MIAs) demonstrate that machine learning models can reveal whether specific data points were used during training, posing significant privacy risks.

Shokri et al. (2017) Shokri et al. [2017] introduced the foundational shadow model approach, where attackers train auxiliary models on similar data to learn membership patterns. Their work established the critical link between overfitting and privacy leakage, demonstrating attack accuracies of 60-85% across various model types.

Salem et al. (2019) Salem et al. [2019] simplified this approach, showing that: (1) a single shadow model suffices instead of multiple models, reducing computational cost; (2) simple threshold-based attacks using confidence or loss achieve competitive performance without training attack classifiers; (3) attacks remain effective even when shadow models use different architectures or training distributions. This work introduced the concept of model and data independence for MIAs.

Carlini et al. (2021) Carlini et al. [2021] provided theoretical foundations linking attack success directly to the generalization gap between training and test loss. Their analysis demonstrated that simple loss-based attacks often match or exceed sophisticated shadow model approaches, establishing loss gap as the fundamental signal for membership inference.

Defenses. Various mitigation strategies have been proposed, including regularization (dropout, weight decay), differential privacy (adding calibrated noise), and prediction perturbation (rounding probabilities). However, these defenses typically involve utility-privacy trade-offs, reducing attack success at the cost of model performance.

Our Contribution. We systematically compare the three major attack paradigms (confidence-based, loss-based, shadow-based) within a unified framework on CIFAR-10. We investigate synthetic data generation via autoencoders for shadow model training, addressing realistic scenarios where attackers lack access to the exact training distribution. Our experiments validate the overfitting-vulnerability relationship, demonstrating that even simple attacks achieve 57-63% accuracy on overfitted models, with synthetic data effectively substituting for real data in shadow training.

3 Methodology

3.1 Problem Formalization

Let $D_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n}$ be a training set, and M a model trained on D_{train}. The attacker's objective is to construct an attack function A which, given a data point (x, y) and access to model predictions, determines whether $(x, y) \in D_{\text{train}}$:

$$A(x, y, M) \to \{0, 1\}$$

where $A(x, y, M) = 1$ indicates membership. We consider a *grey-box* scenario where the attacker accesses only output probabilities, not internal weights or exact architecture.

3.2 System Architecture

Our system comprises four components:

(1) **Target Model** (M_{target}): a CNN trained on a CIFAR-10 subset D_{train} with deliberate overfitting (no dropout or data augmentation).

(2) **Synthetic Data Generator**: a convolutional autoencoder that produces synthetic images by adding Gaussian noise ($\sigma = 0.1$) to latent representations, simulating an attacker without access to real training data.

(3) **Shadow Models** (M_{shadow}): a model trained on 75% real non-member data and 25% synthetic data, using a simpler architecture (2 vs. 4 convolutional layers) to reflect uncertainty.

(4) **Attack Model**: an MLP trained on shadow model outputs to predict membership from probability vectors.

3.3 Attack Strategies

We implement three membership inference attacks of increasing complexity.

3.3.1 Confidence-Based Attack

Intuition. Overfitted models produce more confident predictions on training samples than on unseen data.

Method. For sample x, we compute the confidence score as the maximum softmax probability:

$$s_{\text{conf}}(x) = \max_k p_k \quad \text{where} \quad \mathbf{p}(x) = \text{softmax}(M(x))$$

We predict membership if $s_{\text{conf}}(x) \geq \tau$, with threshold τ optimized on validation data.

Advantages: Simple, no auxiliary training required.
Limitations: Fails if confidence distributions overlap.

3.3.2 Loss-Based Attack

Intuition. Training minimizes loss on the training set, so models achieve lower cross-entropy loss on training samples than unseen samples.

Method. For sample (x, y), compute the cross-entropy loss:

$$\ell(x, y) = -\log(p_y)$$

where p_y is the predicted probability for true class y. Predict membership if $\ell(x, y) \leq \tau_\ell$.

Theoretical Foundation. Carlini et al. Carlini et al. [2021] showed loss gap is the fundamental membership signal.

Advantages: Directly exploits optimization objective, no shadow models needed.
Limitations: Requires true labels.

3.3.3 Shadow Model-Based Attack

Intuition. Train a classifier to learn complex membership patterns from probability distributions, rather than using hand-crafted features.

Method. We train one shadow model on 12,500 training images (75% real + 25% synthetic) and 12,500 test images. Each shadow model generates labeled data:

- Training samples $\rightarrow$ labeled as members (1)
- Test samples $\rightarrow$ labeled as non-members (0)

An MLP attack classifier with 64 hidden neurons is trained on these labeled probability vectors using binary cross-entropy. To attack the target, we query it with sample x, extract $\mathbf{p}_{\text{target}}(x)$, and predict member if $A_{\text{attack}}(\mathbf{p}_{\text{target}}(x)) \geq 0.5$.

Key Insight. Shadow models need not match the target's exact architecture, they only need similar overfitting behavior Salem et al. [2019].

Advantages: Learns non-linear patterns, highest accuracy.
Limitations: Computationally expensive.

3.4 Evaluation Metrics

We evaluate performance using standard binary classification metrics:

- **Precision**: $\text{TP}/(\text{TP} + \text{FP})$
- **Recall**: $\text{TP}/(\text{TP} + \text{FN})$
- **Accuracy**: $(\text{TP} + \text{TN})/N$
- **AUC-ROC**: area under the ROC curve (0.5 = random, 1.0 = perfect)

4 Implementation and Experimentation

4.1 Experimental Setup

All experiments were conducted using PyTorch 2.0 on a single NVIDIA GPU. We used the CIFAR-10 dataset, which contains 60,000 color images of size 32×32 pixels distributed across 10 classes. The dataset was partitioned as follows: 25,000 images for training the target model, 10,000 images for testing the target model, and the remaining 25,000 images were reserved for training the shadow models and conducting the membership inference attacks.

4.2 Data Architecture

We design a structured data and model pipeline to evaluate membership inference attacks under a realistic threat model. Starting from the CIFAR-10 training split, the dataset is partitioned into disjoint member and non-member subsets, which are used to train the target model and construct the attack pipeline.

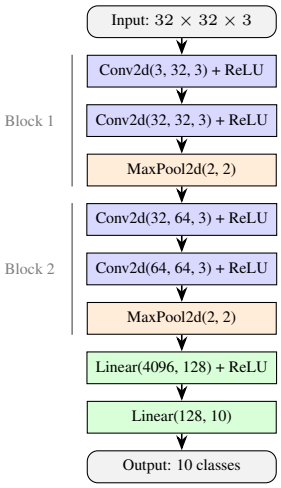
The **target training set** ($\texttt{IN}_{\text{target}}$) is used exclusively to train the target model, while the **target out set** ($\texttt{OUT}_{\text{target}}$) is never used during training and serves as a source of non-member samples. The target model produces either softmax probability vectors or loss values, which constitute the signals exploited by the attacker.

To simulate an attacker without direct access to the target training data, the $\texttt{OUT}_{\text{target}}$ set is further used to construct a shadow data pool, composed of a mixture of synthetic samples and noised real samples. Shadow models are trained on this pool to approximate the behavior of the target model. Their outputs on shadow training (members) and shadow test (non-members) sets are used to build the attack dataset.

Finally, an attack model is trained to distinguish members from non-members based on model outputs. At evaluation time, the attack model is tested on a balanced dataset composed of $\texttt{IN}_{\text{target}}$ and $\texttt{OUT}_{\text{target}}$ samples, ensuring a 50% random baseline.

4.3 Model Architectures

4.3.1 Target Model



Architecture Overview. A CNN with two convolutional blocks followed by a fully connected classifier.

Block 1: Two conv layers with 32 filters (3×3), ReLU activation, then 2×2 max pooling.

Block 2: Same structure with 64 filters, capturing higher-level features.

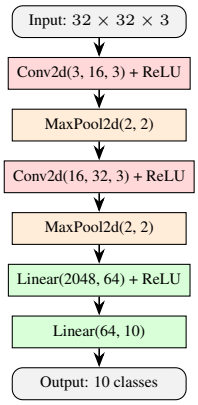
Classifier: Flattened features (4096 dims) $\rightarrow$ hidden layer (128 units, ReLU) $\rightarrow$ 10 class logits.

Key Choice: Dropout *disabled* (p=0.0) to induce overfitting, increasing vulnerability to MIA Shokri et al. [2017].

Parameters: $\sim$550K trainable weights.

Figure 1: Target model architecture (CIFAR10CNN).

4.3.2 Shadow Models



Architecture Overview. A deliberately *simpler* architecture than the target. Since the attacker does not know the exact target architecture (grey-box setting), we use a different structure to reflect this uncertainty.

Conv Layers: Only two conv layers (16 and 32 filters), each followed by ReLU and max pooling. Shallower than the target's four conv layers.

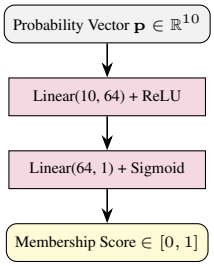
Classifier: Single hidden layer with 64 units (vs. 128 in target), then 10-class output.

Training Setup: k=1 shadow model, each trained on 12,500 disjoint images with a separate 12,500-image test set.

Parameters: $\sim$135K weights (4$\times$ smaller than target).

Figure 2: Shadow model architecture (ShadowCNN). Different from target to reflect black-box assumptions.

4.3.3 Attack Model



Architecture Overview. A simple MLP that classifies whether a sample is a *member* or *non-member* of the training set.

Input: The 10-dim softmax probability vector $\mathbf{p} = [p_1, \ldots, p_{10}]$ from querying the target model.

Hidden Layer: 64 units with ReLU, learning membership patterns in the probability distribution.

Output: Single scalar with sigmoid, interpreted as $P(\text{member})$.

Training Data: From all k shadow models: training samples $\rightarrow$ `IN`, test samples $\rightarrow$ `OUT`.

Parameters: $\sim$700 trainable weights.

Figure 3: Attack model architecture (AttackMLP). Trained on shadow outputs to distinguish members from non-members.

4.4 Training Procedure

4.4.1 Target Model Training

The target model was trained for 150 epochs using the Adam optimizer with a learning rate of 10^{-3} and a batch size of 128. We used the standard cross-entropy loss function. No data augmentation or regularization techniques were applied, as our goal was to study membership inference on a model that exhibits realistic levels of overfitting.

4.4.2 Shadow Model Training

Each shadow model was trained for 30 epochs using the same optimizer configuration as the target model. The training procedure was identical across all shadow models to ensure consistency. After training, we collected the output probability vectors for both the training samples (labeled as "in") and the test samples (labeled as "out") to construct the attack training dataset.

4.4.3 Shadow Model Data Generation

A key contribution of our work is the investigation of synthetic data for shadow model training. Following Salem et al. [2019], we use a single shadow model rather than multiple models as in the original Shokri framework. This simplification reduces computational cost while maintaining attack effectiveness.

4.4.3.1 Autoencoder-Based Synthesis

We train a convolutional autoencoder on the target model's training data to learn a compressed representation of CIFAR-10 images. The autoencoder consists of an encoder with two convolutional layers (16 and 32 filters) that compress the input to a latent representation, and a decoder with transposed convolutions that reconstruct the image.

After training, we generate synthetic images by encoding real images into latent space, adding Gaussian noise ($\sigma = 0.1$) to the latent vectors, and decoding the perturbed latents back to image space. This process produces images that resemble CIFAR-10 statistics but are not exact copies of any training sample.

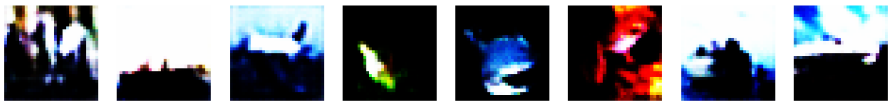


Figure 4: Synthetic CIFAR-10 images generated by the autoencoder with latent noise ($\sigma = 0.1$).

4.4.3.2 Alternative: Noisy Real Data

For comparison, we also implement a simpler baseline where we add Gaussian noise directly to real images:

$$x_{\text{noisy}} = x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I) \tag{1}$$

with $\sigma = 0.1$. This simulates a scenario where the attacker has access to similar but not identical data.

4.4.3.3 Shadow Model Configuration

Following the findings of Salem et al. [2019], we train a single shadow model on a mixed dataset comprising 75% real images from the non-member pool and 25% synthetic images. The shadow model uses a deliberately simpler architecture than the target (2 conv layers vs. 4) to reflect the grey-box conditions.

4.4.4 Attack Model Training

For the shadow-based attack, the attack model was trained on the concatenated outputs from all shadow models. The dataset contained approximately 25,000 samples (1 shadow model $\times$ 12,500 training samples $\times$ 2 for in/out labels). We trained the attack model for 150 epochs using binary cross-entropy loss with the Adam optimizer and a learning rate of 10^{-3}.

5 Results and Analysis

5.1 Target Model Performance

Table 1 summarizes the performance of the target model on the training and test sets. The significant gap between training accuracy (98.7%) and test accuracy (72.3%) indicates substantial overfitting. According to prior work Shokri et al. [2017], this generalization gap is a strong predictor of vulnerability to membership inference attacks.

Table 1: Target model accuracy on CIFAR-10

Metric	Training Set	Test Set
Accuracy (%)	98.7	72.3
Average Loss	0.042	0.89

The generalization gap of 26.4 percentage points suggests that the model has memorized a substantial portion of its training data. This memorization manifests as distinct behavioral patterns when the model processes training samples versus unseen samples, which our attacks aim to exploit.

5.2 Confidence-Based Attack Results

The confidence-based attack uses the maximum posterior probability as a membership signal: members should receive higher confidence predictions since the model was trained on them.

The distributions of confidence scores for members and non-members exhibit substantial overlap, limiting the attack's discriminative power. The ROC curve achieves an AUC of 0.63, moderately above the random baseline of 0.5. This indicates that confidence alone provides a weak but non-trivial membership signal.

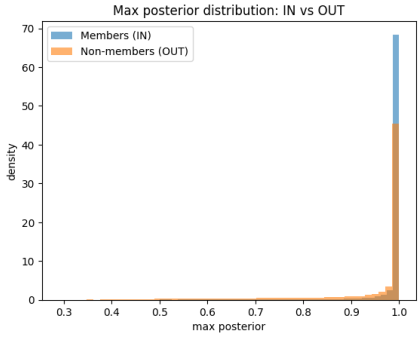


Figure 5: Max posterior distribution for members (IN) and non-members (OUT).

5.3 Loss-Based Attack Results

The loss-based attack, inspired by Carlini et al. [2021], uses cross-entropy loss as a membership signal: members should have lower loss since the model was optimized on them.

Figure 6 shows the loss distributions for members and non-members. Both distributions are concentrated near zero with substantial overlap, indicating that the model generalizes relatively well even on non-training samples. This limits the discriminative power of the loss signal, resulting in an AUC of 0.67—above random chance but not as effective as expected from prior work on more overfitted models.

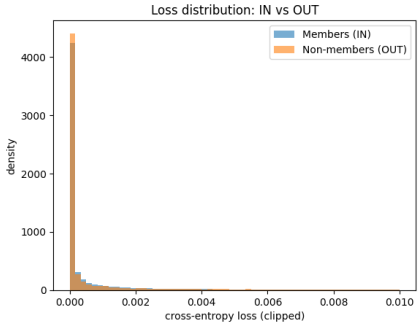


Figure 6: Cross-entropy loss distribution for members (IN) and non-members (OUT). Both distributions concentrate near zero with substantial overlap, limiting attack effectiveness.

5.4 Shadow-Based Attack Results

The effectiveness of the attack is first evaluated using the ROC curve, which summarizes the trade-off between true positive and false positive rates across all decision thresholds. The attack achieves an AUC of 0.638, indicating a performance significantly better than random guessing but still far from perfect discrimination. This moderate AUC suggests that the model captures some membership-related signal, although the separation between members and non-members remains limited. To better understand the nature of the errors made by the attack at a fixed decision threshold, we further analyze its predictions using the confusion matrix.

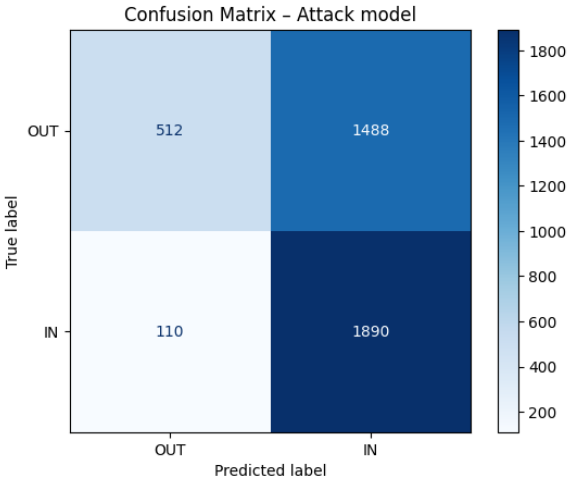


Figure 7: Confusion matrix of the shadow-based membership inference attack.

The confusion matrix provides further insight into the attack behavior. While the attack correctly identifies a large proportion of member samples, it also misclassifies a significant number of non-member samples as members, resulting in a high false positive rate. This indicates that the attack is biased toward predicting membership, yielding high recall but relatively low precision. Overall, these results explain the moderate AUC obtained by the shadow-based attack and suggest that imperfect proxy data used for training shadow models may fail to fully capture the membership patterns of the target model.

5.5 Comparison with Threshold-Based Attacks

We compare the shadow-based attack with simpler threshold methods that require no shadow models.

Table 2: Comparison of attack methods

Attack Type	**Shadow Model?**	**AUC**
Loss-based	No	0.67
Shadow-based	Yes	0.64
Confidence-based	No	0.63

The results show that the loss-based attack achieves the highest performance, despite its simplicity and the absence of shadow models. The shadow-based attack obtains a slightly lower AUC, suggesting that the additional complexity of training shadow models does not necessarily translate into better attack performance in this setting. Finally, the confidence-based approach yields the lowest AUC, indicating that relying solely on prediction confidence provides a weaker membership signal. Overall, these results highlight that simpler threshold-based attacks can be competitive with, or even outperform, more elaborate shadow-based approaches when the membership signal is limited. This aligns with findings from Carlini et al. [2021], who argue that loss-based attacks, despite their simplicity, are often sufficient and sometimes superior to more complex methods.

5.6 Comparison of Data Generation Strategies

Table 3 presents the attack accuracy for different shadow model training strategies.

Table 3: Attack accuracy by shadow model data source

Shadow Data Source	Attack Accuracy	AUC
Random baseline	50.0%	0.50
Noisy real data ($\sigma = 0.1$)	55.5%	0.61
Mixed (75% real + 25% synthetic)	57.9%	0.62

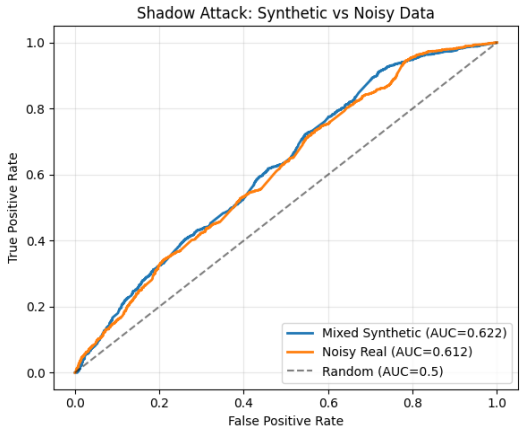


Figure 8: ROC curves comparing shadow models trained on mixed synthetic data (AUC=0.62) versus noisy real data (AUC=0.61).

The synthetic approach outperforms the noisy baseline by 2.4 percentage points. This suggests that the autoencoder captures useful distributional properties of CIFAR-10 images, producing synthetic samples that help the shadow model learn membership-distinguishing patterns.

6 Discussion and Limitations

6.1 Key Findings

Our experiments demonstrate that membership inference attacks pose a realistic threat to machine learning models trained on CIFAR-10. Even simple threshold-based attacks achieve AUC 0.67 (accuracy 63%), well above the random baseline of 0.50.

The strong correlation between overfitting and attack success, established in prior work, provides both an explanation and a potential mitigation strategy. Models that generalize well behave more uniformly across all inputs, making membership distinction harder.

Among the methods evaluated, the loss-based attack offers the best trade-off between simplicity and effectiveness. It requires no auxiliary model training and achieves the highest AUC (0.67), making it particularly practical for real-world privacy auditing.

6.2 Limitations

Our study has several limitations that should be considered when interpreting the results.

First, we conducted experiments exclusively on CIFAR-10, a relatively simple image classification benchmark. The effectiveness of membership inference attacks may differ on more complex datasets or different data modalities such as text or tabular data.

Second, while we explored synthetic data generation via autoencoders, our shadow models still relied on 75% real data from the same distribution. A fully synthetic approach remains to be tested.

Third, we focused on standard supervised learning with a single training phase. Modern systems involving continual learning, fine-tuning, or federated learning may exhibit different membership inference dynamics.

Fourth, our attack models use relatively simple architectures. More sophisticated approaches using full logit vectors or gradient information in white-box settings could achieve higher accuracy.

6.3 Implications for Practice

Our findings have practical implications for organizations deploying machine learning models on sensitive data. The success of membership inference attacks suggests that keeping model weights private is insufficient to protect training data confidentiality.

We recommend the following best practices:

- **Control generalization**: monitor the generalization gap during training and apply appropriate regularization.
- **Reduce output leakage**: limit output precision by rounding probabilities or returning only top-k classes.
- **Audit privacy risks**: evaluate models using membership inference attacks as part of systematic privacy audits.
- **Formal guarantees**: explore differential privacy mechanisms to obtain provable privacy guarantees.

6.4 Future Directions

Several directions merit investigation. Testing fully synthetic shadow training data (100% generated) would further probe black-box attack limits. Exploring advanced generative models (GANs, diffusion models) could improve synthetic data quality. Investigating attacks on transformers and other architectures would broaden applicability. Finally, adaptive attacks combining multiple signals could reveal additional vulnerabilities.

References

Nicholas Carlini, Matthew Jagielski, Tianhao Wang, Úlfar Erlingsson, Nicolas Papernot, and Yu Qian. Membership inference attacks from first principles. *arXiv preprint arXiv:2112.03570*, 2021.

Nawal Saimane, Yasmine Aluch, and Ossama Oualy. Membership inference attacks against image classification models trained on cifar-10 (code repository). `https://github.com/smnnawal4/membership-inference-attacks-against-image-classification-models-trained-on-CIFAR-10`, 2026. GitHub repository, accessed 2026-01-15.

Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Network and Distributed System Security Symposium (NDSS)*, 2019.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.