# ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models

# two main assumptions

- First, the attacker needs to establish multiple shadow models with each one sharing the same structure as the target model. This is achieved by using the same MLaaS that trains the target model to build the shadow models
- Second, the dataset used to train shadow models comes from the same distribution as the target model's training data, this assumption holds for most of the attack's evaluation

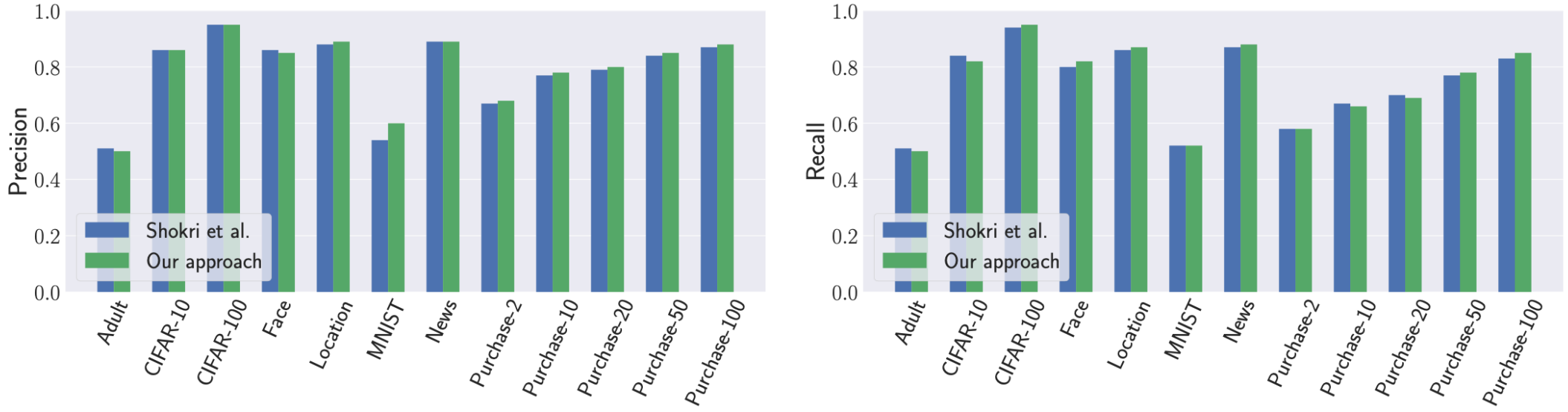
These two assumptions are rather strong which largely reduce the scope of membership inference attacks against ML models

*Adversary 1.* For the first adversary, we assume she has a dataset that comes from the same distribution as the target model's training data. Here, we concentrate on relaxing the assumptions on the shadow models.

- We start by using only one instead of multiple shadow models to mimic the target model's behavior.

- Conclusion：

- Extensive experimental evaluation (we use a suite of eight different datasets ranging from image to text under multiple types of machine learning models) shows that with one shadow model and one attack model, the adversary can achieve a very similar performance as reported by Shokri et al. [38].




(a) Precision.

(b) Recall.

Fig. 1: Comparison of the first adversary's performance with Shokri et al.'s using all datasets. (a) precision, (b) recall.

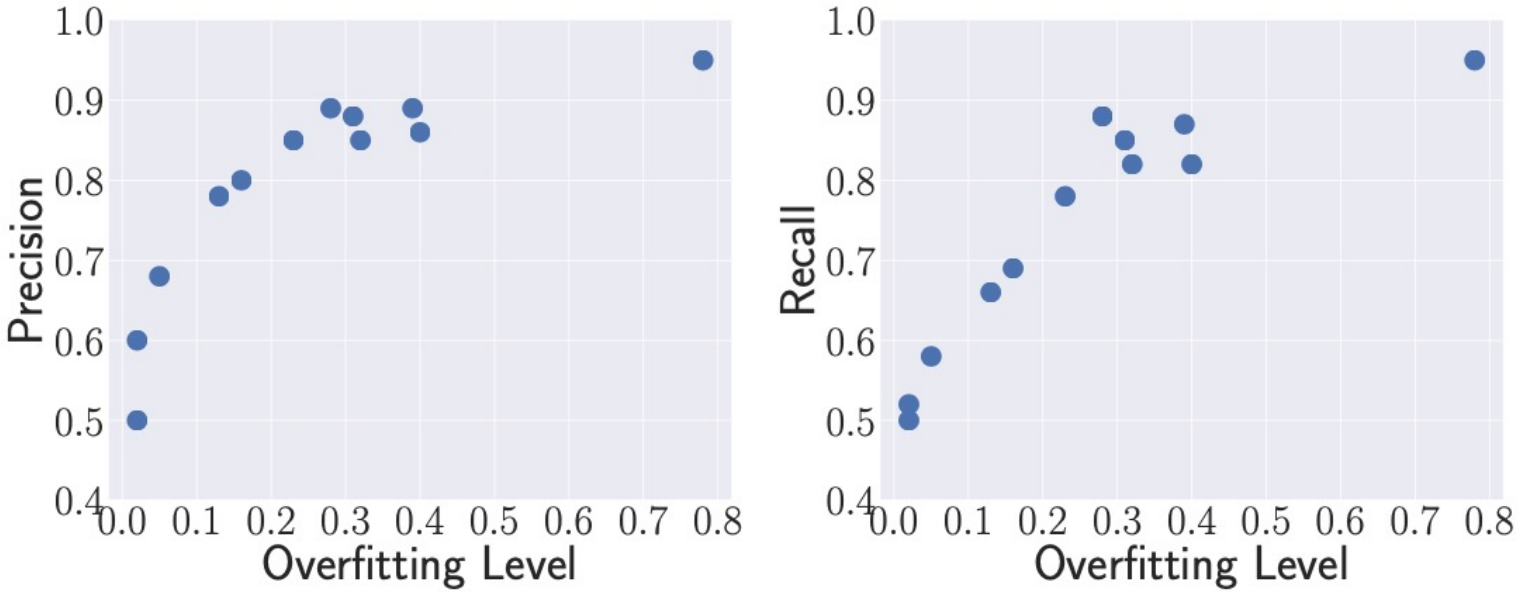



Fig. 2: The relation between the overfitting level of the target model measured by the difference between prediction accuracy on training set and testing set (x-axis) and membership inference attack performance (y-axis). (a) precision, (b) recall.

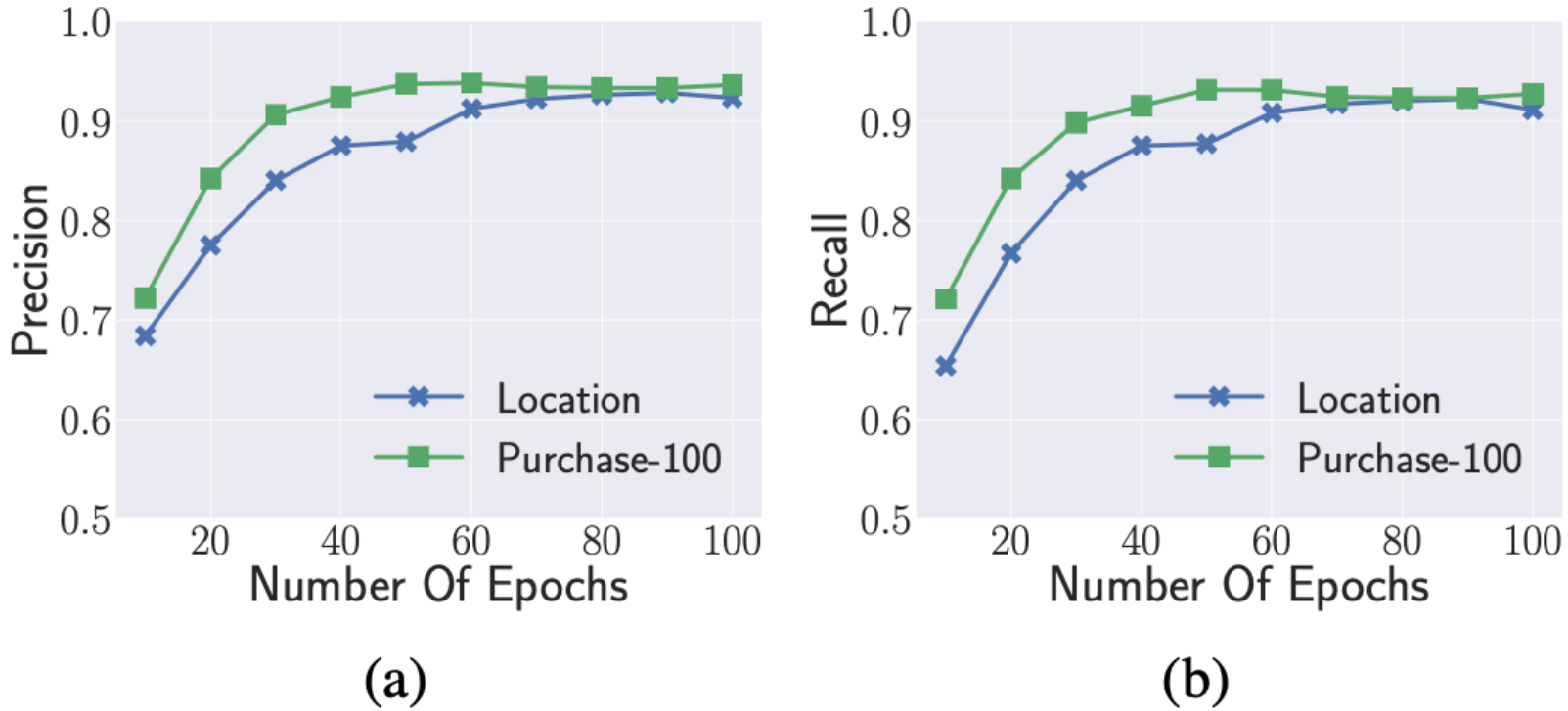
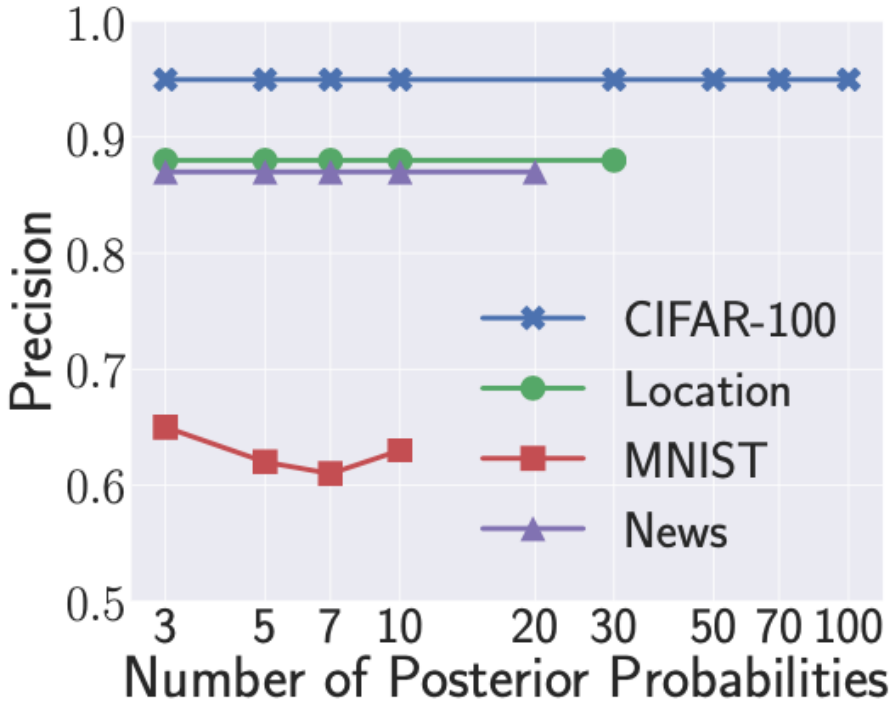
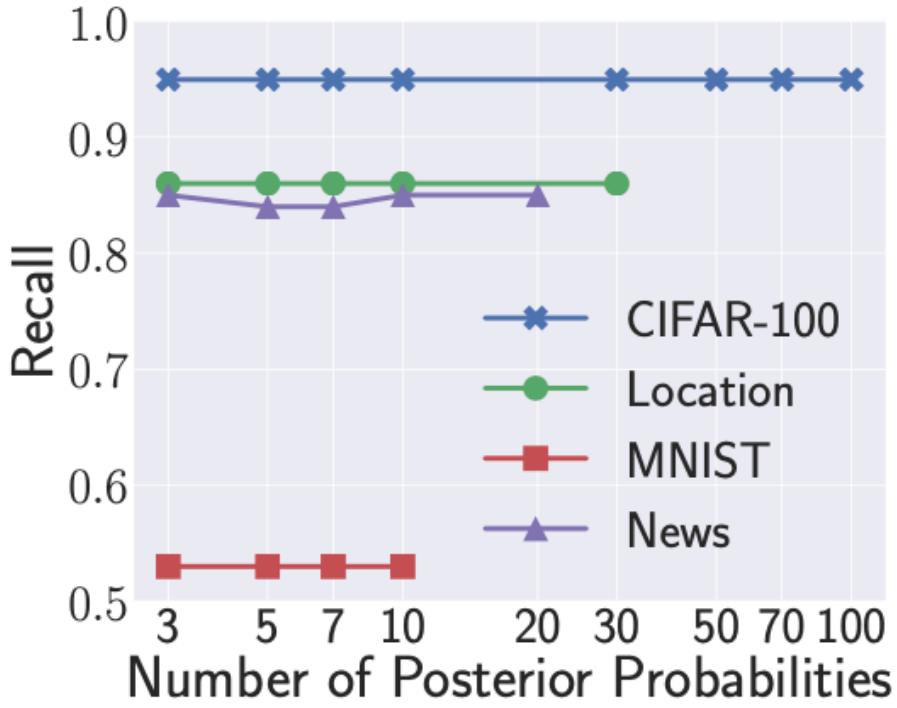



Fig. 3: The relation between the number of epochs used during the training of the target model (x-axis) and membership inference attack performance (y-axis). (a) precision, (b) recall.


1.0
0.9
0.8
0.7
0.6
0.5
Precision
3
5
7
10
20
30
50
70
100
Number of Posterior Probabilities
CIFAR-100
Location
MNIST
News

(a)


1.0
0.9
0.8
0.7
0.6
0.5
Recall
3
5
7
10
20
30
50
70
100
Number of Posterior Probabilities
CIFAR-100
Location
MNIST
News

(b)

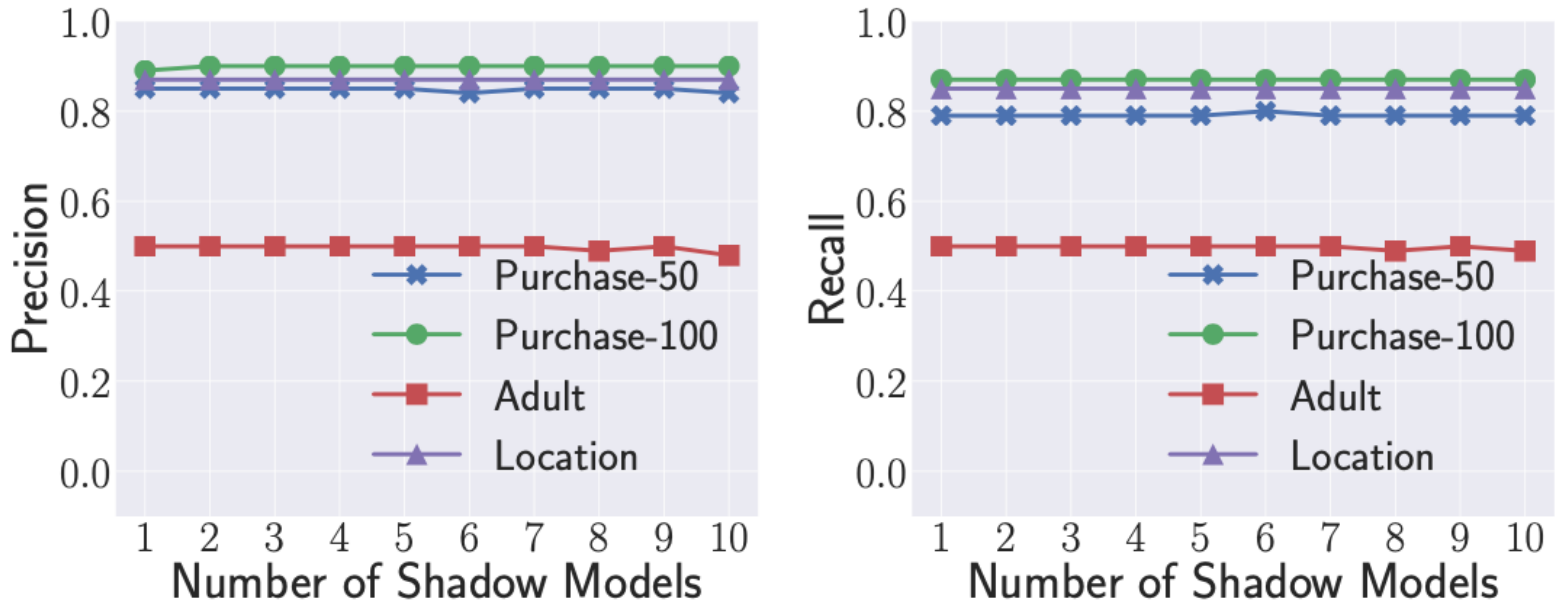



Fig. 5: The effect of the number of shadow models on the first adversary's performance. (a) precision, (b) recall.

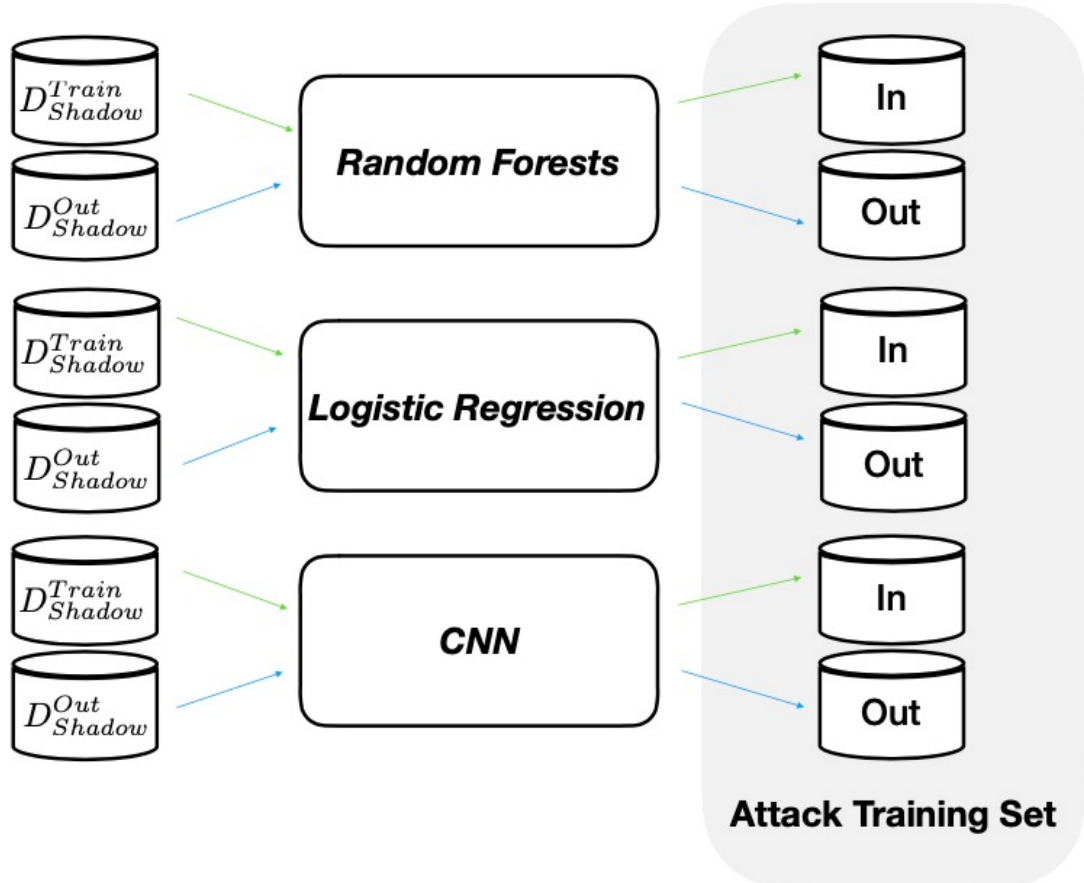



Fig. 6: The architecture of the combining attack on generating data for training the attack model.

| Classifier | With target model structure | | Combining attack | |
|---|---|---|---|---|
| | Precision | Recall | Precision | Recall |
| Multilayer perceptron | 0.86 | 0.86 | 0.88 | 0.85 |
| Logistic regression | 0.90 | 0.88 | 0.90 | 0.88 |
| Random forests | 1.0 | 1.0 | 0.94 | 0.93 |

TABLE II: Comparison of the combining attack and the original attack by the first adversary proposed in Section III-B.

*Adversary 2.* For this adversary, we assume she does not have data coming from the same distribution as the target model's training data. Also, the adversary does not know the structure of the target model. This is a more realistic attack scenario compared to the previous one.

- We propose a *data transferring attack* for membership inference in this setting.

- Conclusion：

- Experimental results show that the membership inference attack still achieves a strong performance, with only a few percentage drop compared to the first adversary.

# *Adversary 3.* This adversary works without any shadow model

- We show that statistical measures, such as maximum and entropy, over the target model's posteriors can very well differentiate member and non-member data points.

- Conclusion：
- Experiments show that such a simple attack can still achieve effective inference over multiple datasets.

Defense. To mitigate the membership risks, we propose two defense mechanisms, i.e., *dropout* and *model stacking*.

*Dropout.* One reason behind membership inference attacks' effectiveness is the inherent overfitting nature of machine learning models.

- Experiments on multiple datasets show that dropout can be a very effective countermeasure against membership inference. On the CIFAR-100 dataset, dropout (with 0.5 dropout ratio) decreases the performance of our first adversary from 0.95 pre- cision and 0.95 recall to 0.61 and 0.60, respectively.
- these models improve in performance *and* resilience in membership inference attacks.
-

*Model Stacking.* Although the dropout mechanism is effective, it is specific to deep neural networks. For target models using other machine learning classifiers, we propose a second defense mechanism, namely model stacking.

- both precision and recall of the attack (adversary 1) drop by more than 30% on the CIFAR- 100 dataset trained with model stacking. Meanwhile, the target model's prediction performance stays almost the same.

- Principle：

- we construct the target model with three different machine learning models. Two models are placed in the first layer directly taking the original training data as input, while the third model is trained with the posteriors of the first two models.