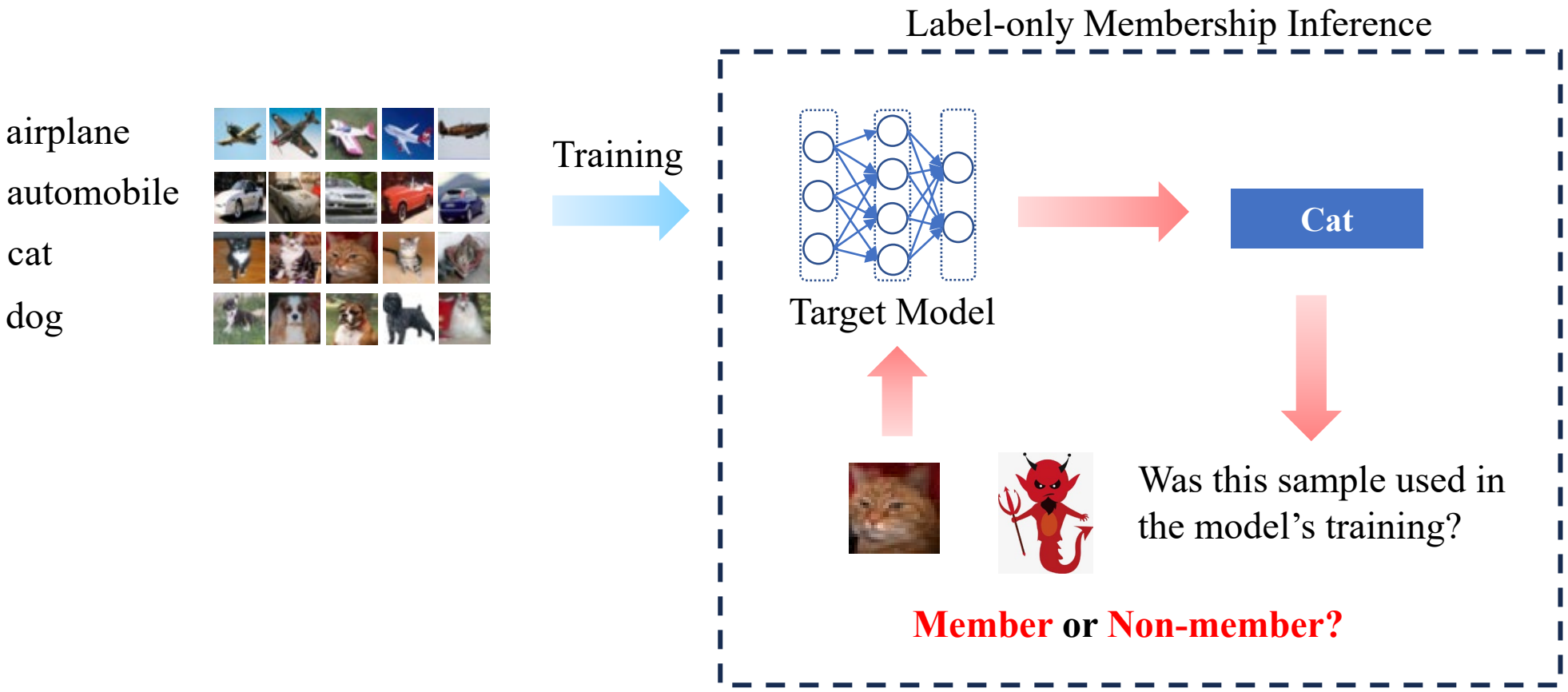
Enhanced Label-Only Membership Inference Attacks with Fewer Queries (USENIX Security2025)

Hao Li*[1], Zheng Li*[2], Siyuan Wu[1], Yutong Ye[1], Min Zhang†[1], Dengguo Feng[1], Yang Zhang[3]

1. Institute of Software, Chinese Academy of Sciences, Beijing, China
2. Shandong University, Shandong, China
3. CISPA Helmholtz Center for Information Security, Saarbrücken, Germany

Source code is available at https://github.com/AIPAG/DHAttack

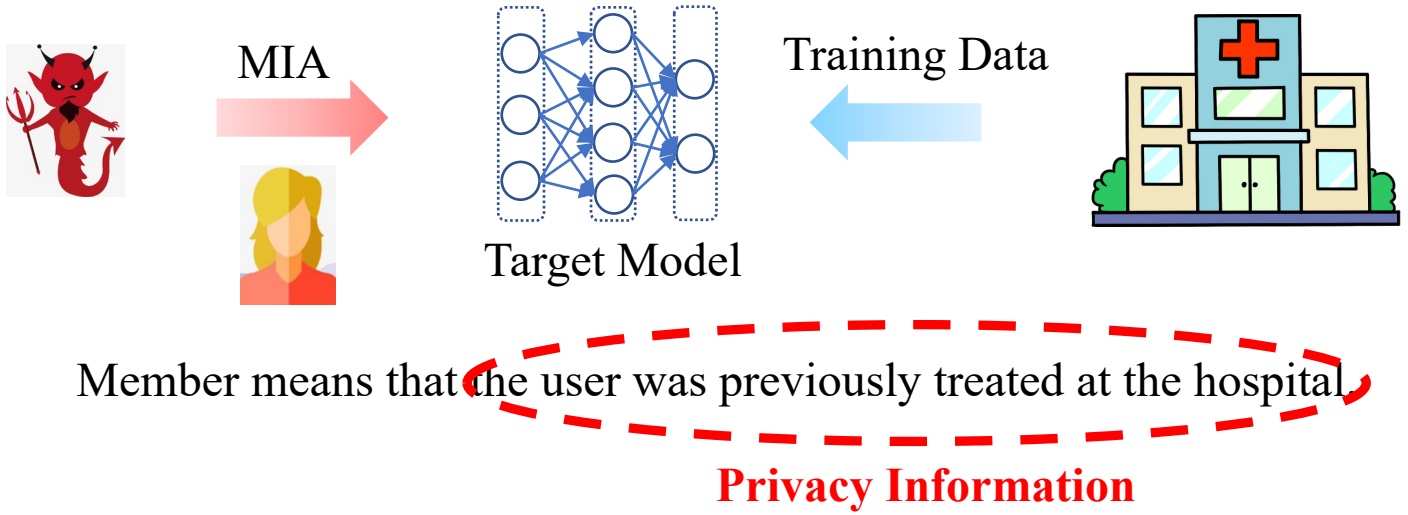
What is label-only membership inference?



The attacker tries to figure out if a specific data sample was used to train a model — but they **only get to see the model's final predicted labels**, nothing else.

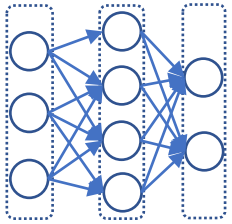
Why we focus on this topic?

- User Privacy Infringement.
- Identifying Potential Privacy Risks in Existing Models.
- Ensuring Compliance with Data Privacy Regulations (e.g., GDPR, CCPA).
- Advancing Privacy Defense Mechanisms.



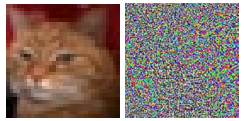
Representative related work

Li et al. 2021(UBA) and Choquette-Choo et al. 2021(SBA)

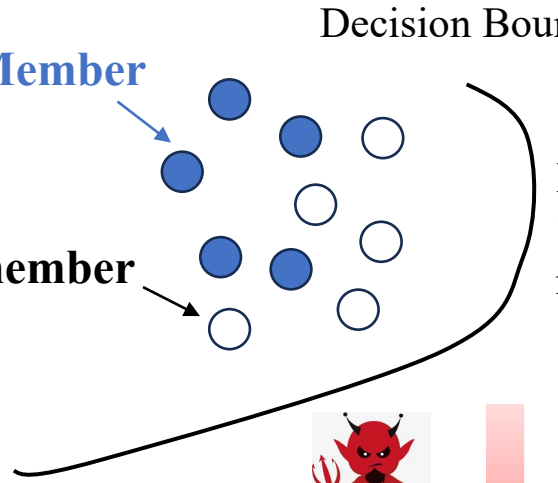


Target Model

From cat to dog or other classes



By adding noise perturbations to the image **thousands of times**, the adversarial example with the **minimal perturbation** (called **ShortestBD**) is found.



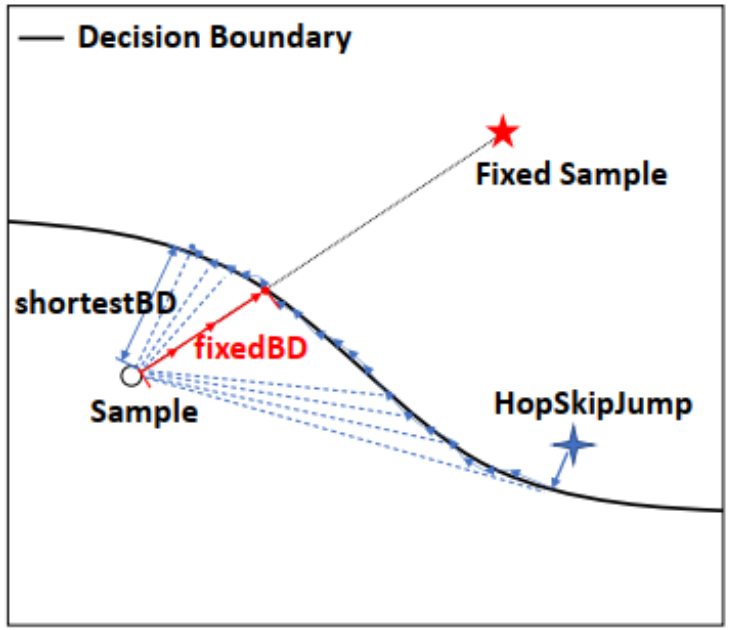
In general, members have larger ShortestBD values compared to non-members.

ShortestBD as the Membership Score

Member or Non-member?

Problems	High query requirement and Low distinguishability

Problem 1： High query requirement

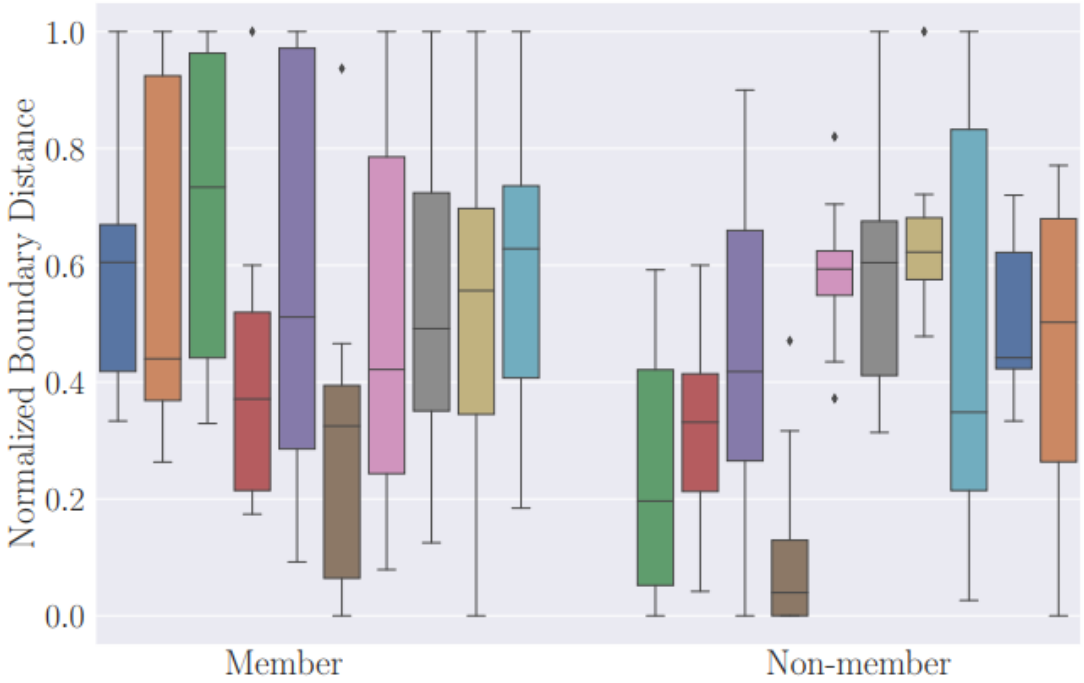


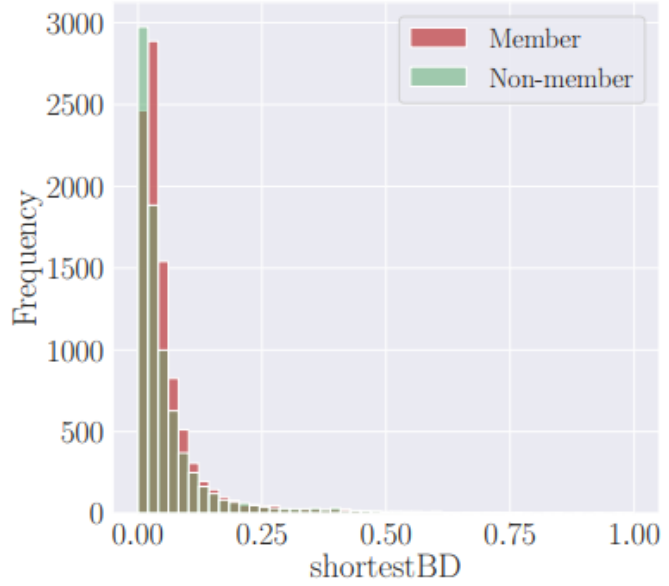
(a) Number of Queries

Existing label-only attacks that employ advanced adversarial example techniques, such as **HopSkipJump** or **QEBA**, typically need **thousands of** queries to identify the shortestBD.

Is such a huge query cost really necessary?

Problem 2： Low distinguishability



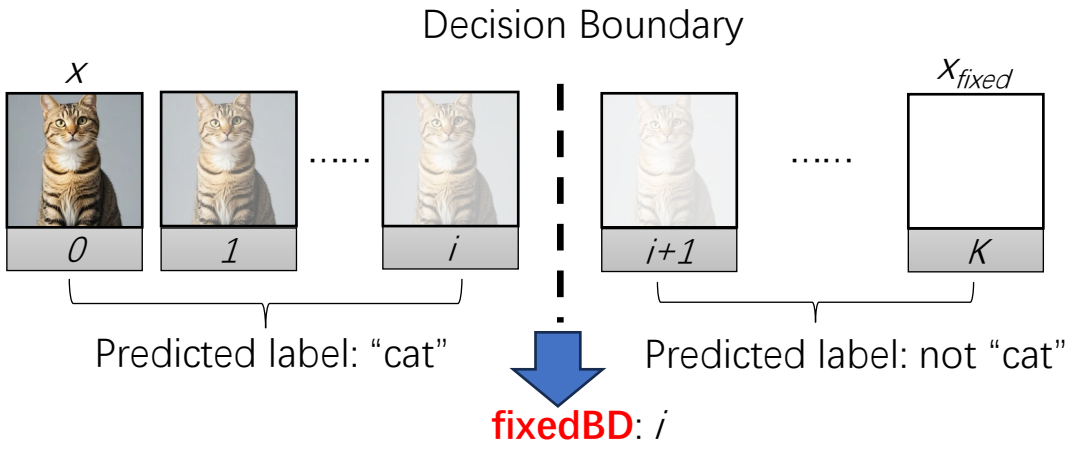


(a) ShortestBD Distribution

Existing label-only MIAs based on adversarial attack algorithms often expend many queries on the target model to identify a suboptimal membership signal, i.e., ShortestBD.

Approach: DHAttack

Computation of fixedBD

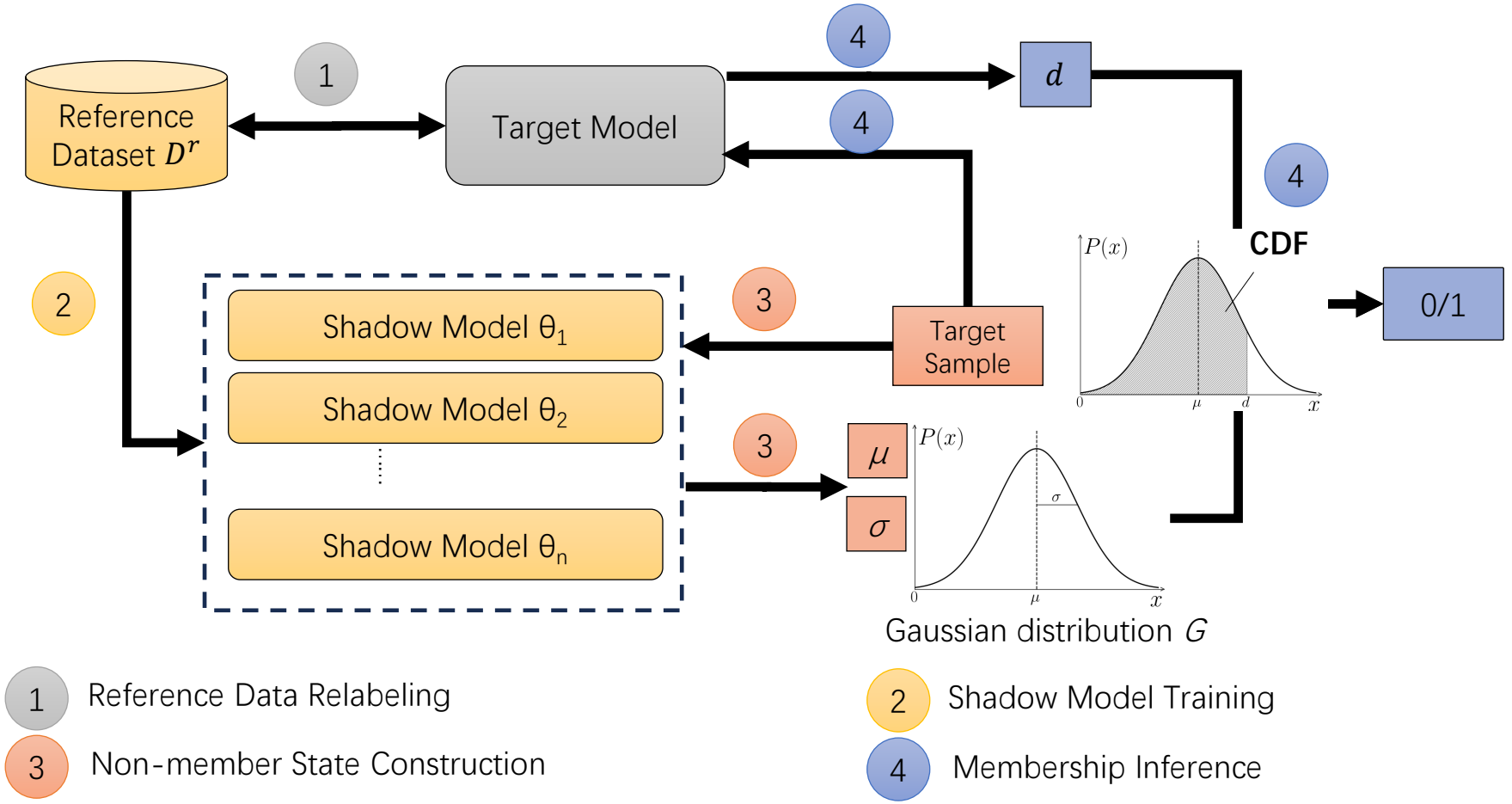


(a) CIFAR10 (b) CIFAR100

We approximate ShortestBD by moving the sample in a fixed direction until it crosses the decision boundary, defining this as fixedBD, and **reduce queries from 2000 to 50**.

Approach: DHAttack

We estimate the fixedBD distribution for non-members via shadow models and Gaussian modeling. Samples exceeding most values in this distribution on the target model are likely members, i.e., **members have a larger CDF value under this distribution**.



Evaluation

Datasets with diverse modalities

CIFAR10, CIFAR100, CINIC10, Purchase, News

Models

VGG-16, ResNet-56, MobileNetV2

Baselines

- Noise Robustness Attack (NRA, Choquette-Choo et al. 2021)
- Unsupervised Boundary-attack (UBA, Li et al. 2021)
- Supervised Boundary-attack (SBA , Choquette-Choo et al. 2021)
- TrajectoryMIA for Label-only (TrajectoryMIA, Liu et al. 2022)
- YOQO (Wu et al. 2023)

Evaluation

Attack performance for image datasets

Table 3: The best attack performance of `DHAttack` and baselines against three target models trained on CIFAR10.

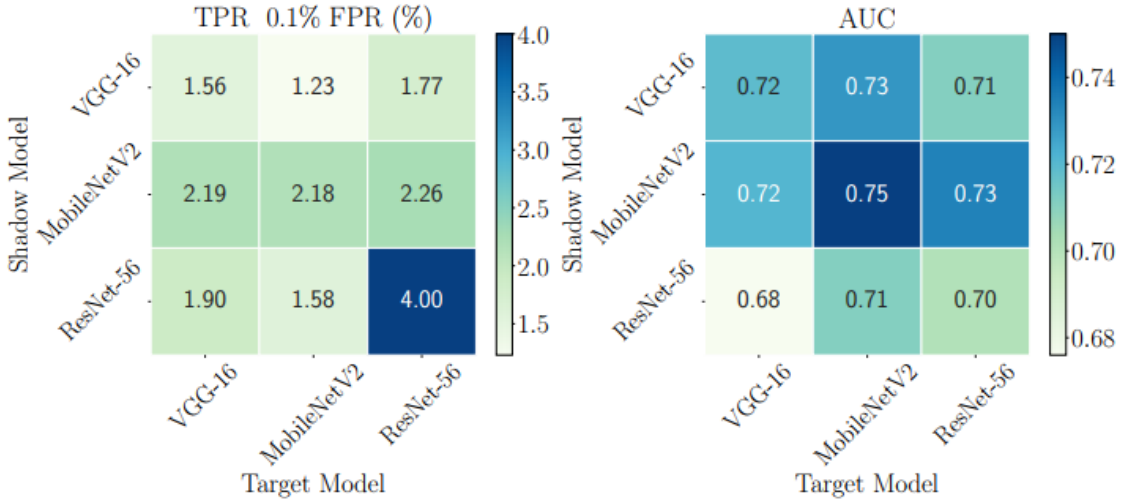
MIA method	TPR @ 0.1% FPR (%)			AUC		
	VGG-16	ResNet-56	MobileNetV2	VGG-16	ResNet-56	MobileNetV2
NRA	0.17(0.3k)	0.14(0.1k)	0.15(1k)	0.700(0.3k)	0.608(0.1k)	0.647(1k)
UBA	0.19(21k)	0.17(0.7k)	0.17(15k)	0.726(21k)	0.605(0.7k)	0.561(15k)
SBA	0.19(6.5k)	0.17(11k)	0.18(11k)	0.725(6.5k)	0.694(11k)	0.702(11k)
TrajectoryMIA	0.34(1k)	0.14(1k)	0.17(1k)	**0.730(1k)**	0.615(1k)	0.642(1k)
YOQO	0.18(1)	0.18(1)	0.17(1)	0.718(1)	0.717(1)	0.696(1)
DHAttack	**1.56(30)**	**2.58(50)**	**2.93(50)**	0.719(30)	**0.752(50)**	**0.750(50)**

DHAttack achieves optimal performance with **only 30–50 queries**, while baselines often require thousands or even tens of thousands. It also outperforms them in most cases, with **TPR at 0.1% FPR** exceeding baselines by **several times, even over 10×**.

Evaluation

Relax Assumptions

Attack performance of DHAttack when the adversary uses the same distribution as the target model's training set versus different distributions.



Attack performance of DHAttack using different model architectures for training shadow models.

Evaluation

Alternative fixedBD Measurements

Different Methods	TPR @ 0.1% FPR (%)			AUC		
	VGG-16	ResNet-56	MobileNetV2	VGG-16	ResNet-56	MobileNetV2
Blurriness	**1.69**	0.90	**3.30**	0.642	0.597	0.653
Rotation	0.87	0.18	1.22	0.653	0.550	0.575
Resize	0.08	0.08	0.30	0.498	0.500	0.390
RGB-0	1.22	**2.53**	**4.04**	**0.719**	**0.753**	**0.756**
RGB-255	**1.56**	**2.58**	2.93	**0.719**	**0.752**	**0.750**

Evaluation

Evaluation of Robustness

MIA method	TPR @ 0.1% FPR (%)				AUC			
	No defense	MixupMMD	DP-SGD	LDL	No defense	MixupMMD	DP-SGD	LDL
NRA	0.18	0.12	0.10	0.14	0.700	0.548	0.511	0.615
UBA	0.19	0.12	0.11	0.15	0.721	0.545	0.501	0.608
SBA	0.19	0.12	0.11	0.14	0.722	0.546	0.520	**0.638**
DHAttack	1.56	**0.22**	**0.20**	**0.54**	0.719	**0.564**	**0.538**	0.620

Conclusion

We propose a new label-only membership inference attack called DHAttack, designed for **Higher performance** and **Higher stealth**.

Higher Performance

- We shift our focus to each individual sample: a sample will exhibit a larger boundary distance if it was in the training set compared to if it itself was not.

Higher Stealth

- We measure the boundary distance to a fixed point, reducing the large query requirements for shortestBD.

Thank you!

For additional information contact us:
lihao@iscas.ac.cn