An Empirical Study of Black-box based Membership Inference Attacks on a Real-world Dataset

Yujeong Kwon, Simon S. Woo, and Hyungjoon Koo

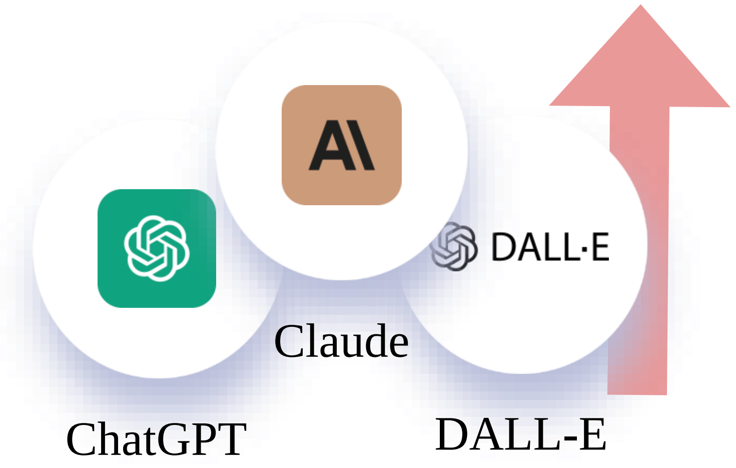
FPS 2024
Montréal, Canada
December 9-11, 2024

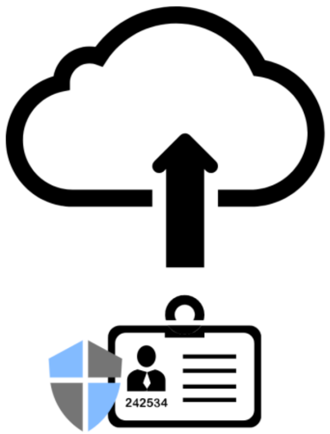
SecAI Lab
SUNGKYUNKWAN UNIVERSITY

Potential Risk in MLaaS

- Data security risks
  - Increasing the use of Machine Learning as a Service (MLaaS) platforms
  - Growing concerns about data security

    → Potential threats to the security of data samples

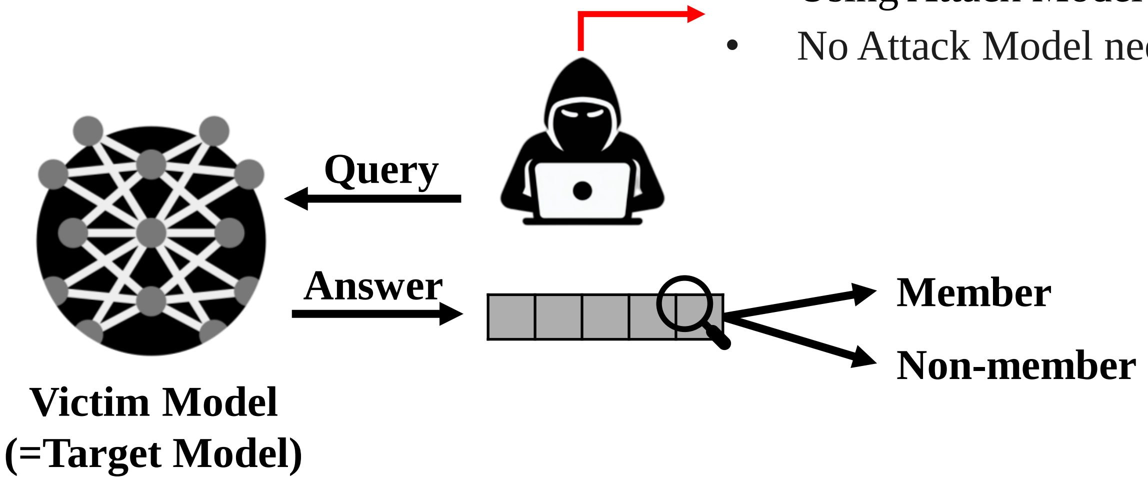




Background

- Membership Inference Attack (MIA)
  - Threatening the security of the data itself
  - Identifying the presence of a specific sample when training a target model

- Using Attack Model → Classifier–based
- No Attack Model needed → Metric-based



Existing MIAs

- Varying techniques depending on the adversary’s knowledge
  - White-box based MIAs
    : Victim model’s architecture, parameters, and distribution of the dataset
    → Requires a strong assumption

  - Black-box based MIAs
    : Part of the knowledge of a white-box adversary
    → Different assumptions
    → Adopt standard (well-known) and (relatively simple) benchmark datasets such as MNIST, CIFAR-10, and CIFAR-100

Black-box based MIAs

- Two types of Black-box MIAs
  - Classifier-based MIAs
    - Use prediction vectors from the victim (or shadow) model
    - Train a binary classifier attack model

  - Metric-based MIAs
    - Use prediction vectors from the victim (or shadow) model
    - Calculate metrics (e.g., correctness, loss, entropy)
    - Compare results with a predefined threshold

Siamese-based MIA

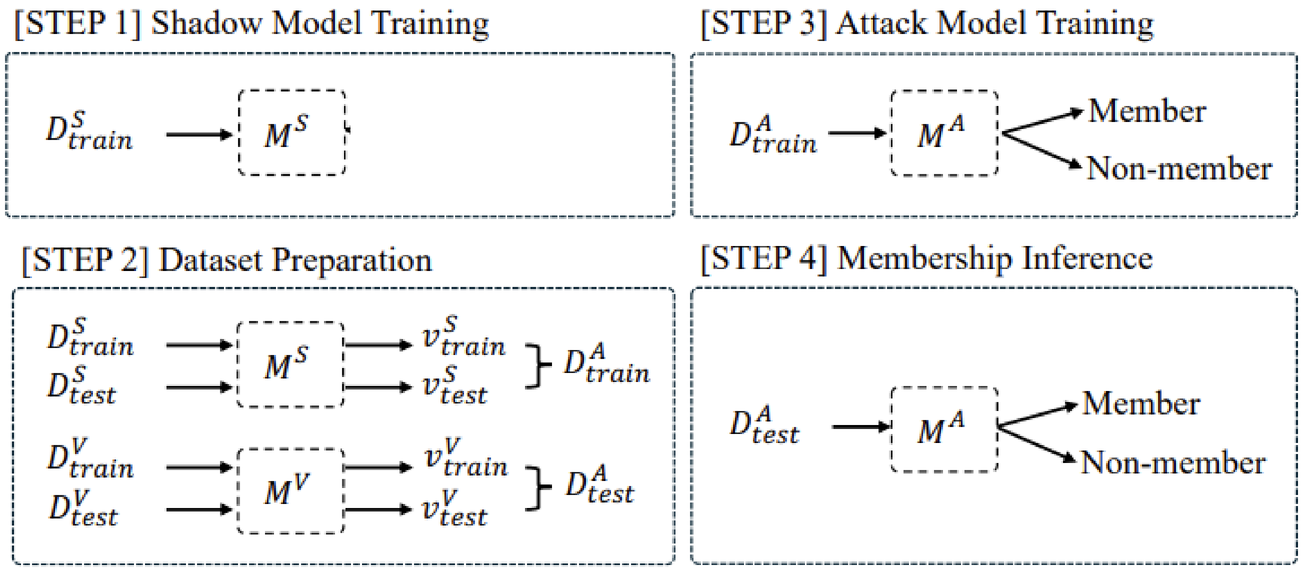
- Attacker's knowledge

  1) Query access to the victim model

  2) Architecture of the victim model

  3) Distribution of the victim model dataset for preparing a shadow dataset

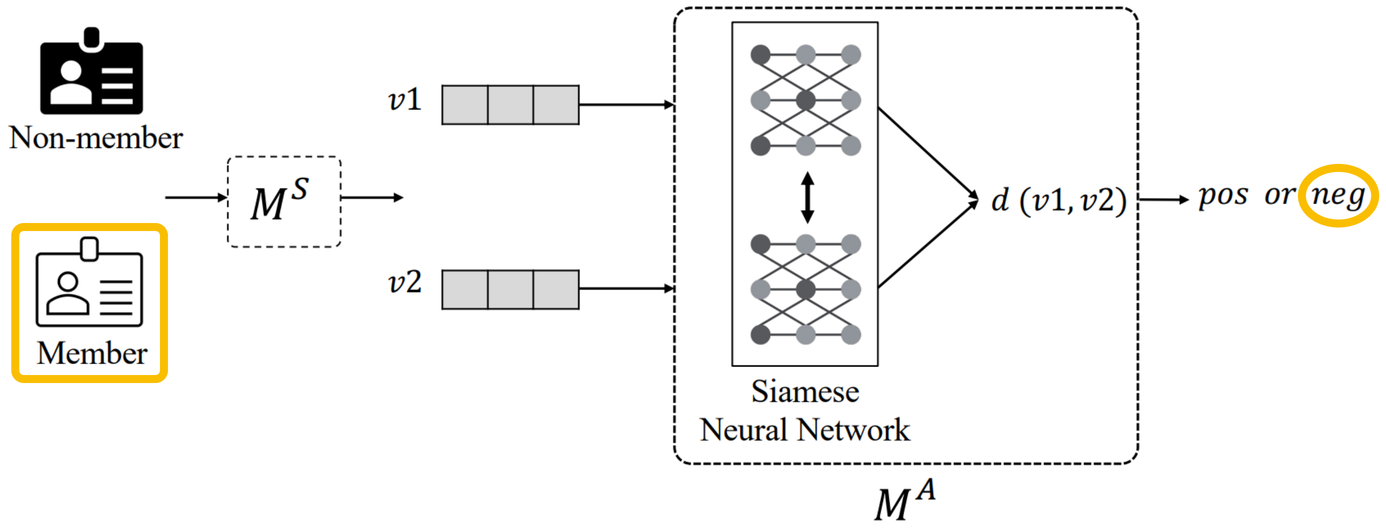
  4) Non-member sample(s)



Overall Workflow

M^S: Shadow Model
M^V: Victim Model
M^A: Attack Model
D^S_{train}: Train dataset of the Shadow model
D^v_{train}: Train dataset of the Victim model
D^A_{train}: Train dataset of the Attack model

Siamese-based MIA



Member – Member : *Positive*
Non-member – Non-member : *Positive*
Member – Non-member : Negative

Research Questions

RQ1. How effective are existing black-box based MIAs and our approach (Siamese-based MIA)

on a previous benchmark dataset (e.g., CIFAR-10)?

RQ2. How well do the MIAs perform against a real-world dataset (e.g., KID34K)?

RQ3. How well the reconstructed images improve MIA performance on a real-world dataset?

Experimental Settings & Results (RQ1)

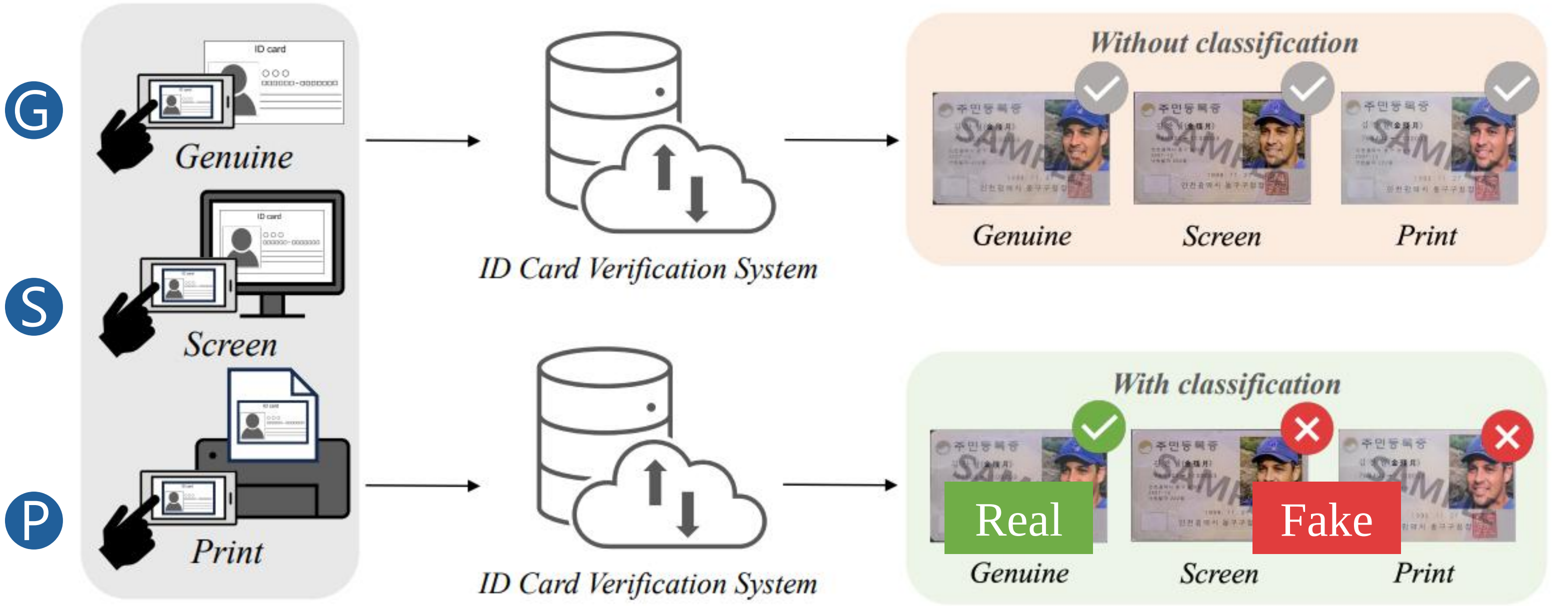
- Dataset: CIFAR-10
- Baselines: 6 metric-based MIAs and 3 classifier-based MIAs
- Our approach: Siamese-based MIA
- Evaluation metric: AUC

Attack Technique	Base Approach	AUC
MIA Evaluation (Threshold attack) [34]	Metric-based (Loss)	0.52
MIA Evaluation (Threshold entropy attack) [34]	Metric-based (Entropy)	0.51
LiRA [5]	Metric-based (Loss)	0.58
Privacy Meter (Population metric attack) [39]	Metric-based (Loss)	0.50
Privacy Meter (Shadow metric attack) [33,39]	Metric-based (Correctness)	0.50
Privacy Meter (Reference metric attack) [39]	Metric-based (Loss)	0.50
MIA Evaluation (Logistic regression attack) [34]	Classifier-based	0.51
SAMIA [40]	Classifier-based	0.67
Confidence-based neural network attack [33]	Classifier-based	0.64
Siamese-based MIA (Ours)	Classifier-based	0.70

Effectiveness of MIA against a Real-world Dataset

- MIA against a target model on the real-world dataset
    - Images that contain sensitive information
    - High-resolution images (512x800 ≈ 409.6K)

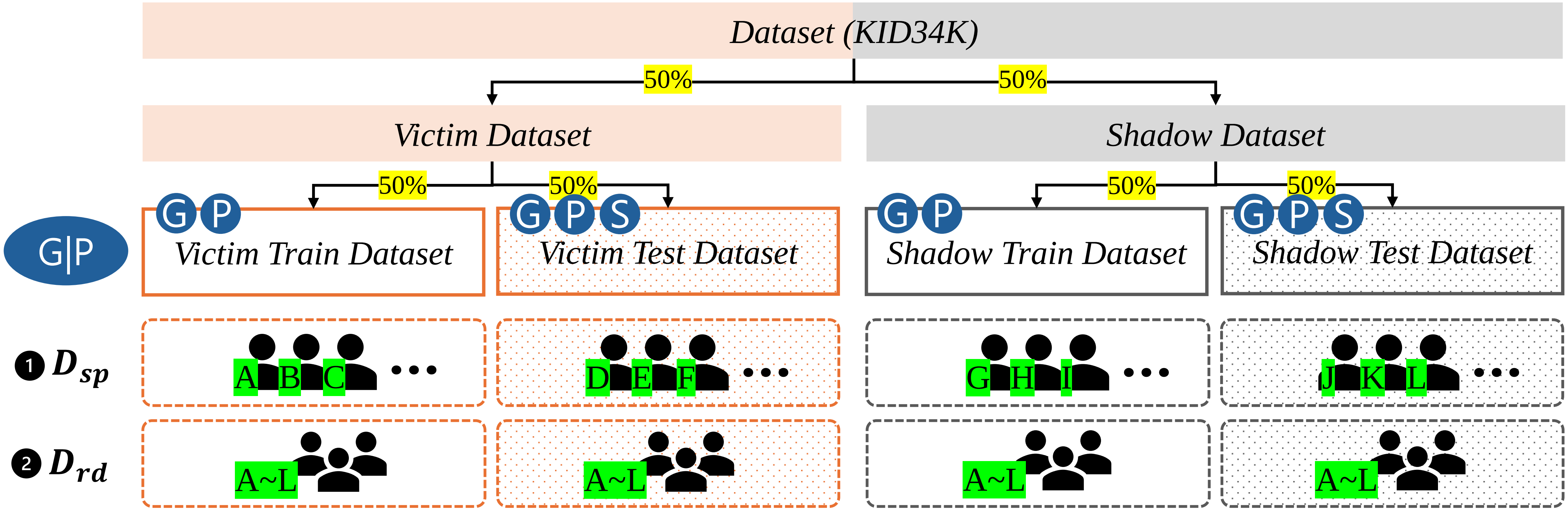
→ KID34K (512x800) dataset: ID cards and drivers' licenses



Park, Eun-Ju, et al. "KID34K: A Dataset for Online Identity Card Fraud Detection." *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2023.

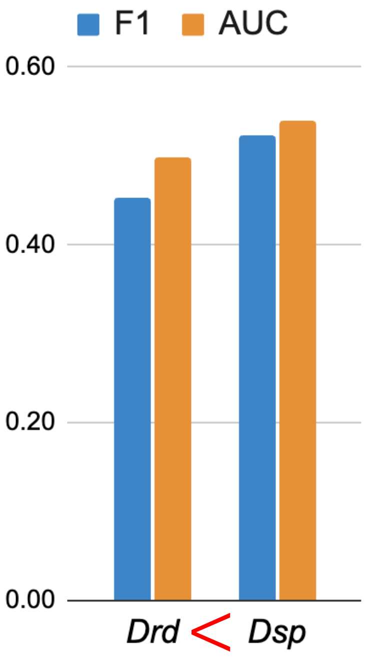
Various Dataset Compositions for MIA Performance

- User-based
  - ❶ D_{sp} (splits by individual user): A realistic assumption for checking the membership of a single person
  - ❷ D_{rd} (randomly splits each dataset)
- Label-based
  - G: Genuine, P: Print, S: Screen

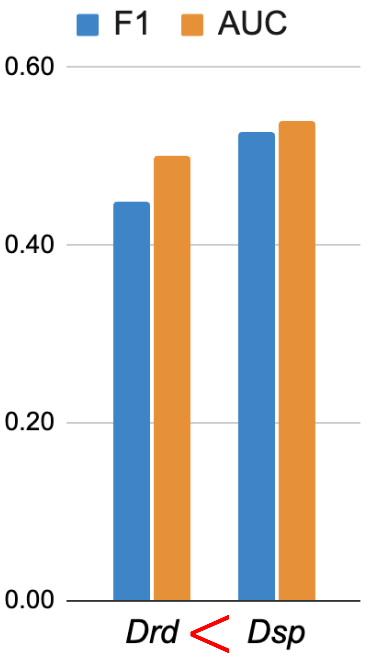


Experimental Settings & Results (RQ2)

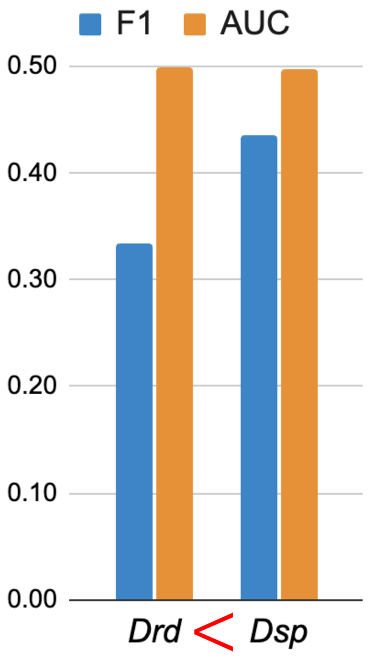
- Dataset: KID34K
- Baselines: 2 classifier-based MIAs (selected from the best performance in RQ1)
- Evaluation metrics: F1, AUC



SAMIA



Confidence-based



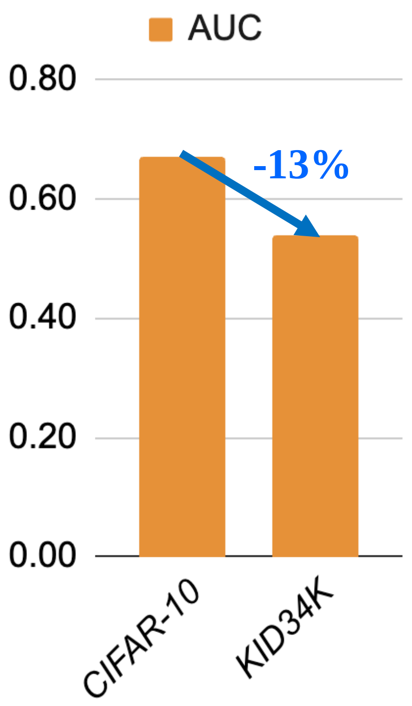
Siamese-based (Ours)

$D_{rd} < D_{sp}$

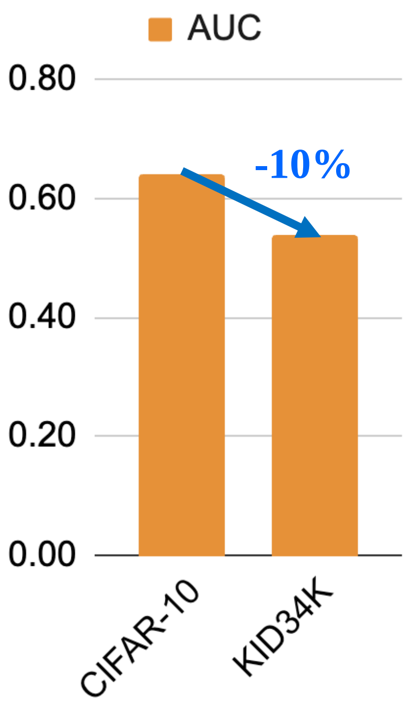
- D_{rd} (randomly splits each dataset)
- D_{sp} (splits by individual user)

Experimental Settings & Results (RQ2)

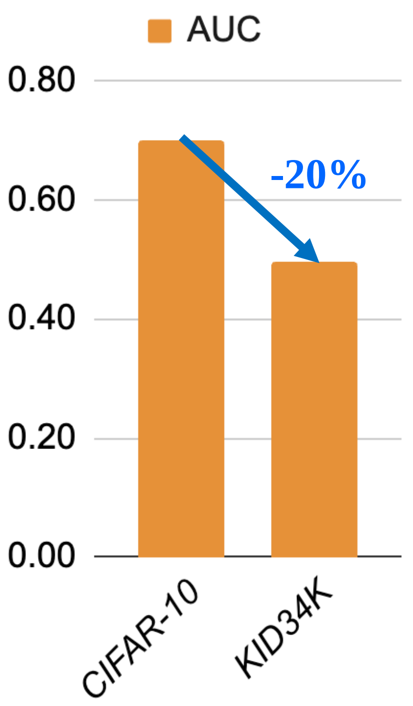
- Dataset: KID34K
- Baselines: 2 classifier-based MIAs (selected from the best performance in RQ1)
- Evaluation metrics: AUC



SAMIA



Confidence-based



Siamese-based (Ours)

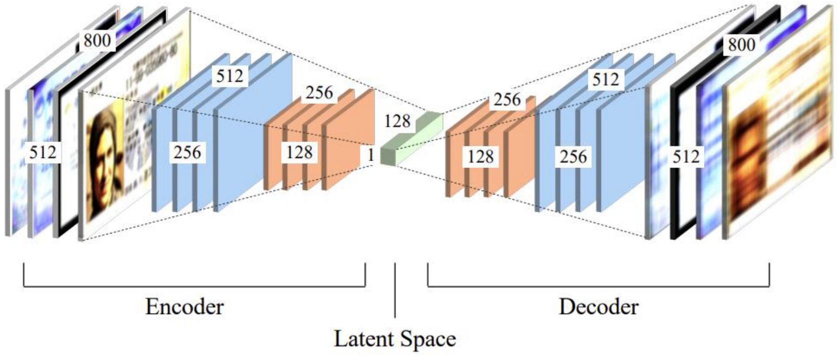
Our Idea: Reconstructed Images

- Performance of MIA techniques degrades with excessive features

Dataset	Resolution
CIFAR-10	32 x 32
KID34K	512 x 800

- MIA configuration and sample properties may affect the accuracy of membership inference

? How can we improve MIA performance on a real-world dataset?
→ Reducing resolution meaningfully can improve MIA performance
→ Reconstructing images with an autoencoder



Impact of Reconstructed Images on MIA Performance (RQ3)

- Images reconstructed by an autoencoder help in training a shadow model

D_{train}	*Genuine*	*Print*	*Screen*
$G\|P$	v	v	
$G\|S$	v		v
All	v	v	v

Selection	D_{train}	Method	F1	AUC
D_{sp}	$G\|P$	SAMIA	$\boxed{0.56}$	$\boxed{0.62}$
D_{sp}	$G\|S$	SAMIA	0.53	0.53
D_{sp}	*All*	SAMIA	0.52	**0.54**
D_{sp}	$G\|P$	Confidence-based	**0.63**	**0.67**
D_{sp}	$G\|S$	Confidence-based	0.54	0.54
D_{sp}	*All*	Confidence-based	0.53	**0.54**
D_{sp}	$G\|P$	Siamese-based*	0.45	0.51
D_{sp}	$G\|S$	Siamese-based*	0.41	0.47
D_{sp}	*All*	Siamese-based*	0.44	**0.50**

Original

Selection	D_{train}	Method	F1	AUC
D_{sp}	$G\|P$	SAMIA	$\boxed{\mathbf{0.64}}$	$\boxed{\mathbf{0.66}}$
D_{sp}	$G\|S$	SAMIA	0.53	**0.54**
D_{sp}	*All*	SAMIA	**0.53**	0.53
D_{sp}	$G\|P$	Confidence-based	0.62	0.66
D_{sp}	$G\|S$	Confidence-based	0.54	0.54
D_{sp}	*All*	Confidence-based	0.53	0.53
D_{sp}	$G\|P$	Siamese-based*	**0.51**	**0.52**
D_{sp}	$G\|S$	Siamese-based*	**0.42**	**0.48**
D_{sp}	*All*	Siamese-based*	**0.45**	0.49

Reconstructed

Summary of Our Findings

- MIA results can vary depending on
  - Number of features (dimension)
  - Dataset configuration (sample characteristics)

  → Leading to inconsistencies with other datasets

- Autoencoder-generated images enhance the success rate of MIAs
  - 16% performance drop in shadow models by adopting an autoencoder

- Defending against MIAs involves trade-offs between model performance and security

Threats to Validity

- Generalization
  - Limited applicability to diverse datasets (e.g., financial, healthcare)
- Scope
  - White-box MIAs' results may vary

Conclusion

- Black-box MIAs may underperform on a real-world dataset (KID34K)
- Proposing a Siamese-based MIA
- Reducing features can empirically improve MIA performance

Thank you