APPENDIX

A. DATASETS DESCRIPTION

CIFAR100 (C100). CIFAR100 is a widely used benchmark dataset to image classification, which consists of 60,000 images in 100 classes. We randomly select two disjoint sets of 10,000 images as the target models and the shadow models' training datasets, respectively.

CIFAR10 (C10). CIFAR10 is a widely used dataset to evaluate the image classification, which consists of 60,000 images in 10 classes. We randomly select two disjoint sets of 10,000 images as the target models and the shadow models' training datasets, respectively.

CH_MNIST (CH_M). CH_MNIST is a benchmark dataset of histological images used to evaluate human colorectal cancer, consisting of 5,000 histological images in 8 classes of tissues. We follow the same image processing methods and classification tasks as prior BlindMI-DIFF [7] to resize all images to 64×64. We randomly select two disjoint sets of 2,500 images as the target models and the shadow models' training datasets, respectively.

ImageNet (ImaN). Tiny-imagenet is a widely used dataset to image classification, which is a subset of the ImageNet dataset and consists of 100,000 images in 200 classes. We randomly select two disjoint sets of 10,000 images as the target models and the shadow models' training datasets, respectively.

Location30 (L30). Location30 contains location "check-in" records of individuals. We obtain a pre-processed dataset from [4] which contains 5,010 data with 446 binary features corresponding to whether an individual has visited a particular location. All data samples are clustered into 30 classes representing different geosocial types. The classification task is to predict the geosocial type based on the 466 binary features. Following [29], we use 1,000 data to train a target model.

Purchase100 (P100). Purchase100 contains shopping records of different individuals. We obtain a pre-processed dataset from [4] containing 197,324 data samples with 600 binary features corresponding to a specific product. All data samples are clustered into 100 classes representing different purchase styles. The classification task is to predict the purchase style based on the 600 binary features. We follow *Nasr et al.* [30] to use 10% data samples (19,732) to train a target model.

Texas100 (T100). Texas100 consists of Texas Department of State Health Services' information about patients discharged from public hospitals. Each data record contains information about the injury, diagnosis, the procedures the patient underwent and some demographic details. We obtain preprocessed dataset from [4] which contains 100 classes of patient's procedures consisting 67,330 data with 6,170 binary features. The classification task is to predict the patient's main procedure based on the patient's information. Following [29], [30] ,we use 10,000 data samples to train a target model.

B. ORIGINAL AND CONSTRUCTED DISTANCE DISTRIBUTIONS OF DATA SAMPLES IN THE TARGET DATASET

Figure 6 and Figure 8 show the original and constructed distance distribution of data samples in the target datasets.

C. THE SELECTED DISTANCES OF THE TARGET DATASET

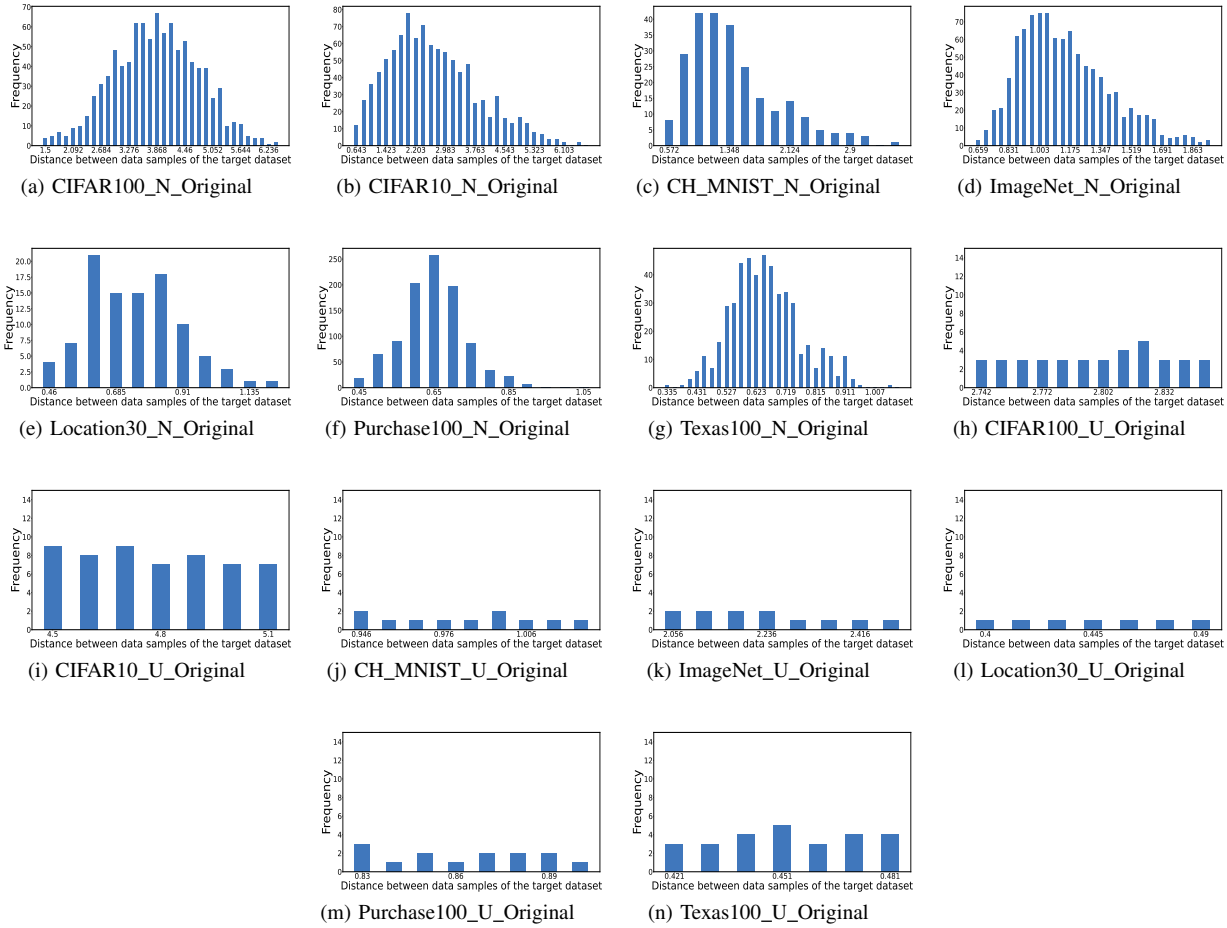
Table VIII shows the selected distances between data samples and differential distances between two datasets in this work.

TABLE VII: The Parameter (p) and thresholds of the Bernoulli Distribution of Data Samples in the Target Dataset.

Parameter (p)	Thresholds of Bernoulli Distribution in Different Datasets						
	C_100	**C_10**	**CH**	**Image**	**L_30**	**P_100**	**T_100**
0.1	2.623	1.147	0.729	0.835	0.520	0.504	0.512
0.2	2.97	1.529	0.837	0.906	0.574	0.554	0.551
0.3	3.306	1.84	0.935	0.964	0.608	0.566	0.578
0.4	3.535	2.097	1.081	1.021	0.660	0.608	0.605
0.5	3.778	2.39	1.186	1.081	0.705	0.620	0.630
0.6	4.013	2.707	1.30	1.153	0.750	0.635	0.661
0.7	4.259	3.064	1.45	1.227	0.784	0.675	0.692
0.8	4.555	3.513	1.765	1.323	0.845	0.688	0.729
0.9	4.913	4.285	2.19	1.491	0.920	0.741	0.813

D. CCRS BETWEEN TWO ATTACKS WHEN TWO OF THE FOUR INFLUENCING FACTORS VARY

Table XI shows the Conflicting Comparison Results (CCRs) between two attacks when two of the four influencing factors vary.



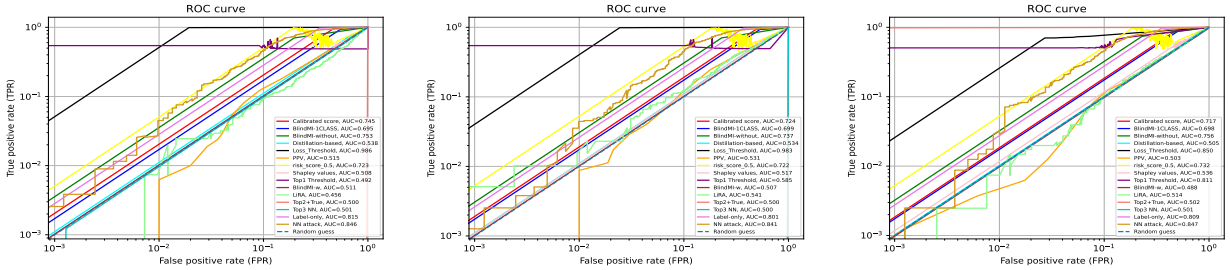
(a) CIFAR100_N_Original (b) CIFAR10_N_Original (c) CH_MNIST_N_Original (d) ImageNet_N_Original

(e) Location30_N_Original (f) Purchase100_N_Original (g) Texas100_N_Original (h) CIFAR100_U_Original

(i) CIFAR10_U_Original (j) CH_MNIST_U_Original (k) ImageNet_U_Original (l) Location30_U_Original

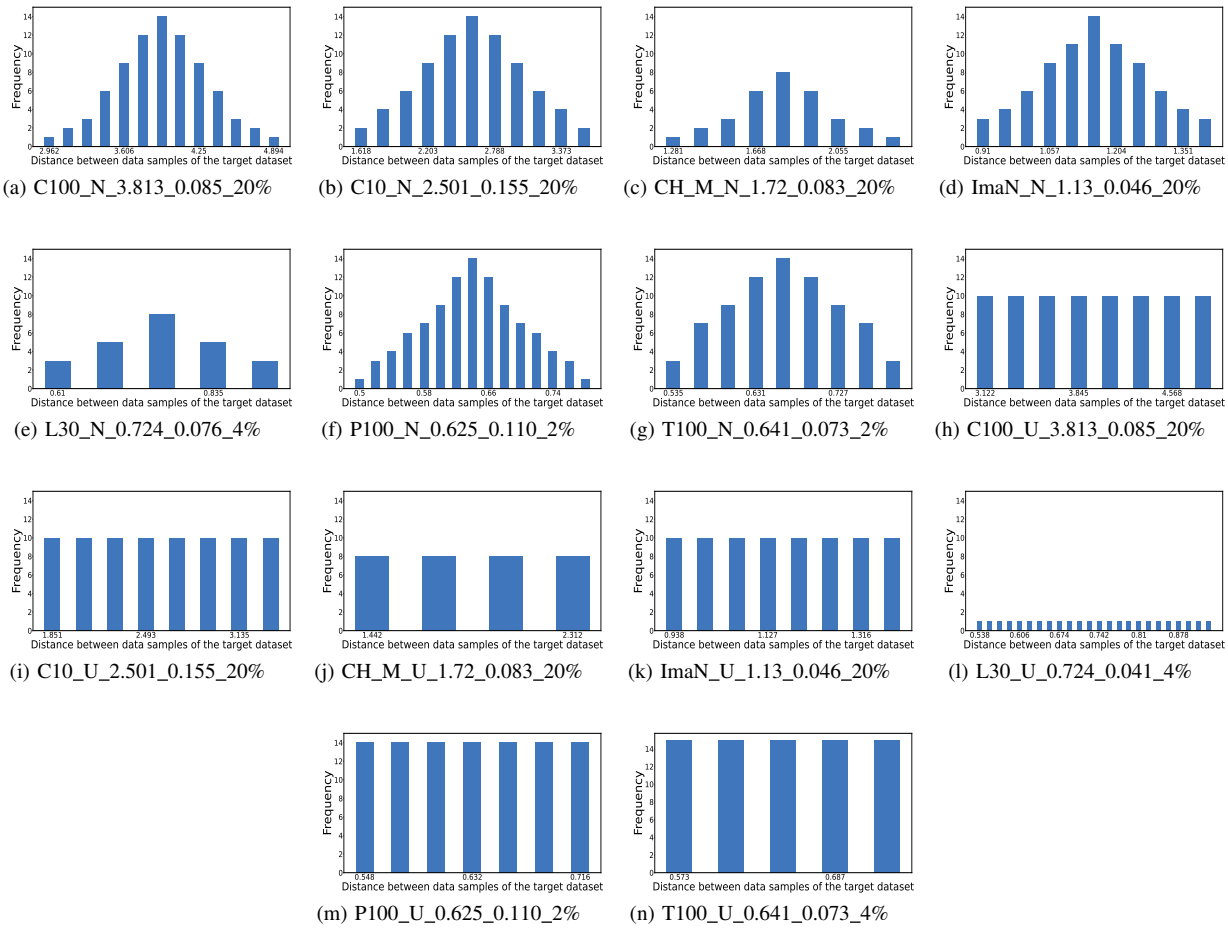
(m) Purchase100_U_Original (n) Texas100_U_Original

Fig. 6: The Original Distance Distribution of Data Samples in the Target Dataset. Inside (a)-(g) are the original distance distributions of data samples in the target CIFAR100, CIFAR10, CH_MNIST, ImageNet, Location30, Purchase100 and Texas100 datasets approximately obey normal distributions, and (h)-(n) are the original distance distributions of data samples in the target seven datasets approximately obey uniform distributions.



(a) C100_N_2.893_0.085_20% (b) C100_U_2.893_0.085_20% (c) C100_B_2.893_0.085_20%

Fig. 7: Comparing the true positive rate vs. false positive rate of prior 15 state-of-the-art membership inference attacks on CIFAR100 by the AUC computed as the Area Under the Curve of attack ROC in logarithmic scale. Inside (a)-(c) are the ROCs in logarithmic scale of the constructed distance distributions of data samples in the target CIFAR100 (C100) dataset obey normal, uniform and bernoulli distributions, respectively.



(a) C100_N_3.813_0.085_20% (b) C10_N_2.501_0.155_20% (c) CH_M_N_1.72_0.083_20% (d) ImaN_N_1.13_0.046_20%

(e) L30_N_0.724_0.076_4% (f) P100_N_0.625_0.110_2% (g) T100_N_0.641_0.073_2% (h) C100_U_3.813_0.085_20%

(i) C10_U_2.501_0.155_20% (j) CH_M_U_1.72_0.083_20% (k) ImaN_U_1.13_0.046_20% (l) L30_U_0.724_0.041_4%

(m) P100_U_0.625_0.110_2% (n) T100_U_0.641_0.073_4%

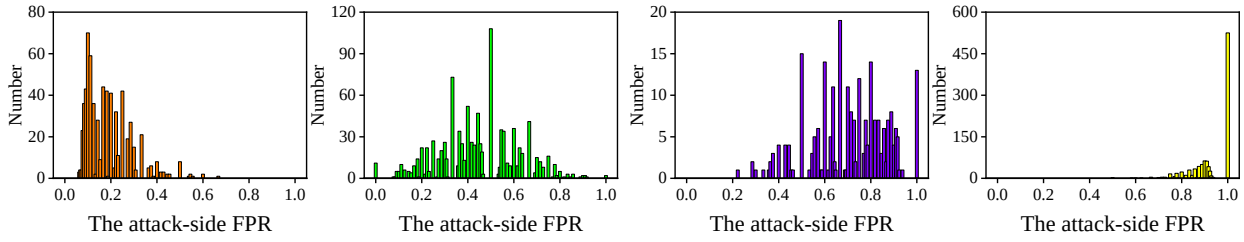
Fig. 8: The Constructed Distance Distribution of Data Samples in the Target Dataset obeys Normal and Uniform Distribution, respectively. Inside (a)-(g) are the constructed distance distributions of data in the target C100, C10, CH_M, ImaN, L30, P100 and T100 datasets obey normal distributions, and (h)-(n) are the constructed distance distributions of data obey uniform distributions.

TABLE VIII: The Selected Distances Between Data Samples (Dis_Between_Data) and Differential Distances Between Two Datasets of the Target Dataset.

Dataset	Dis_Between_Data	Differential Distances
CIFAR100	2.893	0.085
	3.813	0.119
	4.325	0.157
CIFAR10	1.908	0.155
	2.501	0.213
	3.472	0.291
CH_MNIST	0.954	0.083
	1.355	0.108
	1.720	0.133
ImageNet	0.934	0.046
	1.130	0.080
	1.388	0.145
Location30	0.570	0.041
	0.724	0.076
	0.801	0.094
Purchase100	0.550	0.087
	0.625	0.110
	0.729	0.156
Texas1000	0.530	0.038
	0.641	0.073
	0.734	0.107

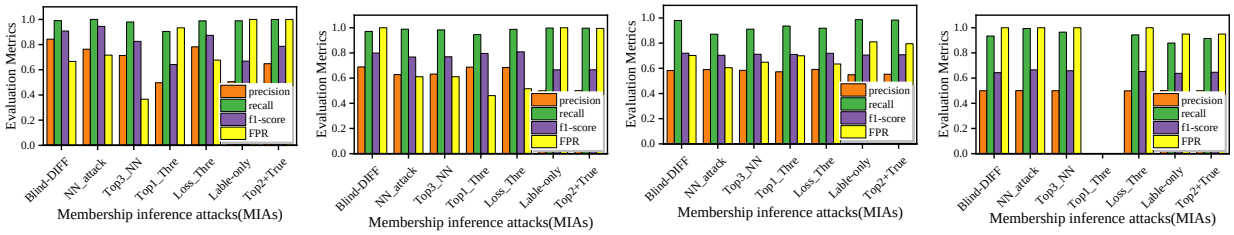
TABLE IX: The effect of the distance between data samples in the target dataset (DisData) on the attacker-side MA.

Dataset	DisData	Attacker-side Membership Advantage (MA)								
		BlindMI-Diff	NN_attack	Label-only	Loss-Thres	Top3-NN	Top2+True	PPV	Calibrated Score	Distillation-based Thre
CIFAR100	**3.823**	**61.63%**	**56.13%**	**75.38%**	**73.37%**	**60.25%**	**69.25%**	**1.83%**	**35.60%**	**25.86%**
CIFAR10	2.573	53.34%	38.50%	33.50%	50.50%	38.38%	46.13%	-0.07%	33.61%	24.56%
CH_MNIST	1.315	27.72%	26.60%	17.60%	20.13%	26.24%	18.84%	-0.92%	22.04%	23.45%
ImageNet	1.138	0.47%	0.15%	0.56%	0.05%	0.08%	0.09%	-1.08%	-0.07%	23.12%



(a) CIFAR100 (b) CIFAR10 (c) CH_MNIST (d) ImageNet

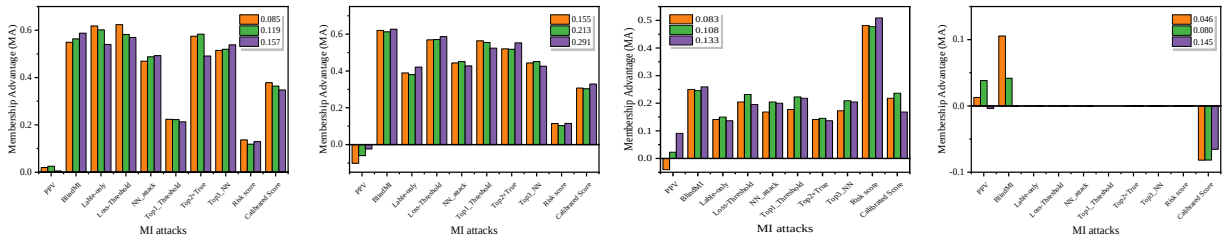
Fig. 9: The distribution of the FPR of the BlindMI-Diff-w/ [7].



(a) CIFAR100 (b) CIFAR10 (c) CH_MNIST (d) ImageNet

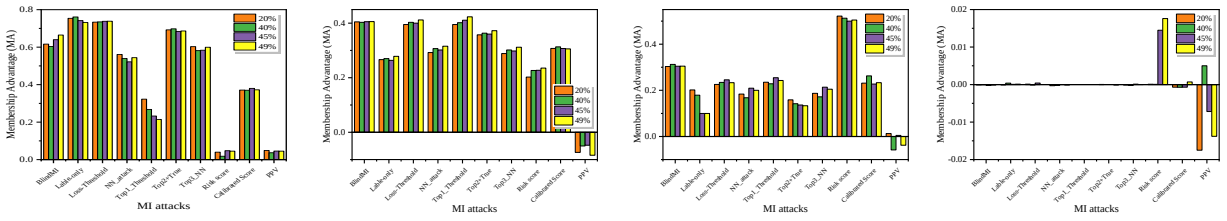
Fig. 10: The evaluation of the existing MI attacks (e.g., *the attacker-side precision, recall, f1-score and FPR*).

.



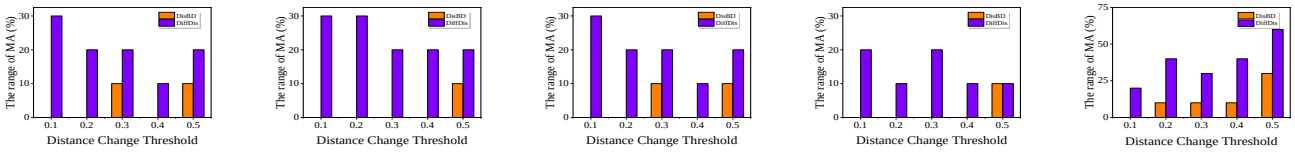
(a) CIFAR100 (b) CIFAR10 (c) CH_MNIST (d) ImageNet

Fig. 11: The effect of differential distance between two datasets on the Membership Advantage.



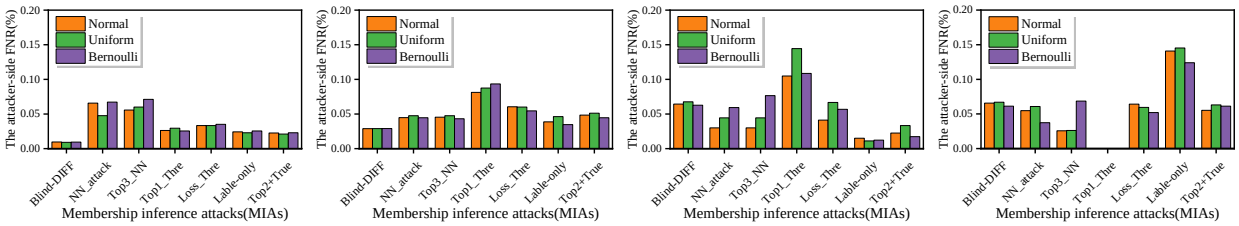
(a) CIFAR100 (b) CIFAR10 (c) CH_MNIST (d) ImageNet

Fig. 12: The effect of the ratio of the samples that are made no inferences by an MI attack.



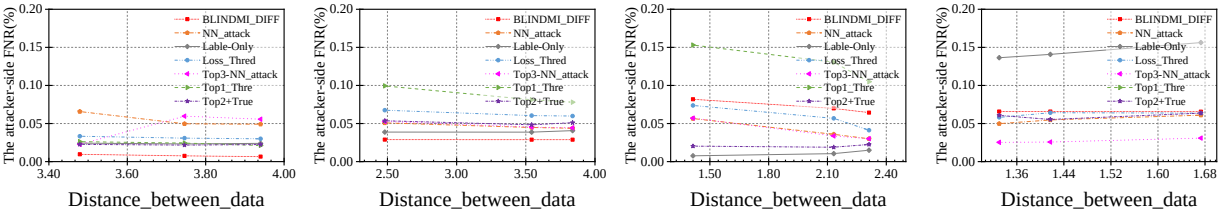
(a) CIFAR100 (b) CIFAR10 (c) CH_MNIST (d) ImageNet (e) Purchase100

Fig. 13: The comparisions of the effect of the distance between data samples of the target dataset (DisBD) and the differential difference between two datasets (DiffDis) on MA.



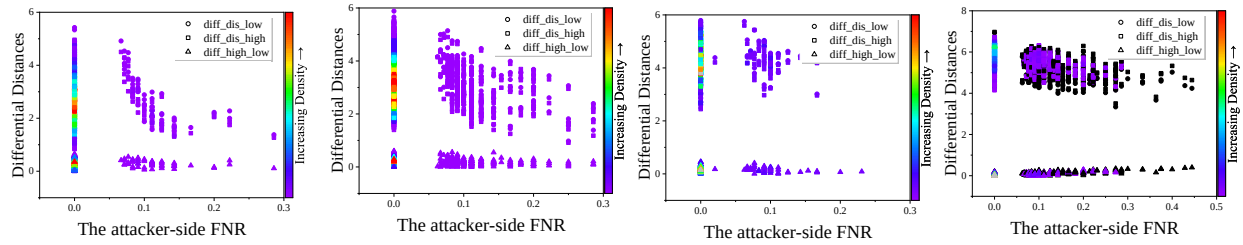
(a) CIFAR100 (b) CIFAR10 (c) CH_MNIST (d) ImageNet

Fig. 14: The attacker-side FNR of the distance distribution of the target datasets obeying normal, uniform and bernoulli distributions.



(a) CIFAR100 (b) CIFAR10 (c) CH_MNIST (d) ImageNet

Fig. 15: The effect of the distance between data samples in the target dataset on the attacker-side FNR.



(a) CIFAR100 (b) CIFAR10 (c) CH_MNIST (d) ImageNet

Fig. 16: The effect of the differential distance between two datasets on attacker-side FNR.

TABLE X: The average distance between data samples (DisBD) and the differential distance between two datasets (DiffDis) change on the MA (C_MA).

Dataset	C_MA	average DisBD	average DiffDis
CIFAR100	10%	0.143	**0.107**
	20%	0.230	**0.099**
	30%	0.349	**0.138**
	40%	0.460	**0.143**
	50%	0.630	**0.257**
CIFAR10	10%	0.113	**0.082**
	20%	0.179	**0.115**
	30%	0.299	**0.182**
	40%	0.440	**0.238**
	50%	0.544	**0.320**
	60%	0.659	**0.389**
	70%	0.793	**0.545**
CH_MNIST	10%	0.140	**0.091**
	20%	0.249	**0.158**
	30%	0.335	**0.237**
	40%	0.476	**0.328**
	50%	0.638	**0.408**
	60%	0.843	**0.547**
ImageNet	10%	0.268	**0.225**
	20%	0.580	**0.288**
	30%	0.805	**0.379**

TABLE XI: Conflicting Comparison Results Between Two Attacks when Two of the Four Influencing Factors Vary.

NumCS.	Attack A	Attack B	PrWk.	Dataset	AtkEff.	Evaluation Scenarios	KeyIF.	NumCS.	Attack A	Attack B	PrWk.	Dataset	AtkEff.	Evaluation Scenarios	KeyIF.
CS01	Top3_NN	PPV	-	4+6	◐ ◑	ES44:P100_U+0.625+0.110+12% ES33:P100_U+0.550+0.110+10%	IF2/IF4	CS25	PPV	NN_attack	7	4	◐ ◑	ES60:ImaN_B+0.934+0.080+40% ES69:ImaN_B+1.130+0.080+20%	IF2/IF4
CS02	Top3_NN	LiRA	-	4+6	◐ ◑	ES23:ImaN_N+1.388+0.046+45% ES12:ImaN_N+1.130+0.046+49%	IF2/IF4	CS26	PPV	BlindMI-without	-	3+4+5	◐ ◑	ES16:ImaN_N+1.130+0.080+49% ES17:ImaN_N+1.130+0.145+20%	IF3/IF4
CS03	Top3_NN	Shapley values	-	3+4+5+7	◐ ◑	ES50:CHM_U+1.720+0.083+40% ES31:CHM_U+0.954+0.083+45%	IF2/IF4	CS27	PPV	Distillation-based	1	1+2+3+4	◐ ◑	ES16:C100_N+3.813+0.119+49% ES24:C100_N+4.325+0.119+40%	IF2/IF4
CS04	Top2+True	Risk score	-	2+3+6	◐ ◑	ES37:P100_U+0.625+0.087+02% ES51:P100_U+0.729+0.087+10%	IF2/IF4	CS28	LiRA	PPV	1	1+2+3+4+5+6	◐ ◑	ES48:P100_U+0.625+0.156+12% ES55:P100_U+0.729+0.156+10%	IF2/IF4
CS05	BlindMI-w	PPV	-	4+5+6+7	◐ ◑	ES45:P100_U+0.625+0.156+02% ES43:P100_U+0.625+0.110+10%	IF3/IF4	CS29	LiRA	NN_attack	1	4	◐ ◑	ES63:ImaN_B+0.934+0.145+45% ES73:ImaN_B+1.130+0.145+20%	IF2/IF4
CS06	BlindMI-w	LiRA	-	4+5+6+7	◐ ◑	ES18:LA30_N+0.724+0.094+08% ES16:LA30_U+0.724+0.076+16%	IF3/IF4	CS30	LiRA	Shapley values	-	2+3+5	◐ ◑	ES38:LA30_U+0.724+0.041+08% ES31:LA30_U+0.570+0.041+12%	IF2/IF4
CS07	BlindMI-w	Top3-NN	3	1+2+3+4+6	◐ ◑	ES48:CHM_U+1.355+0.133+49% ES54:CHM_U+1.720+0.108+49%	IF2/IF3	CS31	LiRA	Calibrated Score	1	5+6+7	◐ ◑	ES12:P100_N+0.625+0.087+12% ES23:P100_N+0.729+0.087+10%	IF2/IF4
CS08	BlindMI-w	Label-only	3	2+3+4	◐ ◑	ES56:CHM_U+1.720+0.133+49% ES34:CHM_U+0.954+0.108+49%	IF2/IF3	CS32	LiRA	BlindMI-without	-	4+5	◐ ◑	ES26:ImaN_N+1.388+0.080+49% ES27:ImaN_N+1.388+0.145+45%	IF3/IF4
CS09	BlindMI-w	Top2+True	3	1+2+3+4	◐ ◑	ES47:C100_U+3.813+0.157+45% ES53:C100_U+4.325+0.119+45%	IF2/IF3	CS33	LiRA	Distillation-based	1	1+2+3+4	◐ ◑	ES20:ImaN_N+1.130+0.145+49% ES27:ImaN_U+1.388+0.145+45%	IF2/IF4
CS10	BlindMI-w	NN_attack	3	2+3+4	◐ ◑	ES48:CHM_U+1.355+0.133+49% ES54:CHM_U+1.720+0.108+49%	IF2/IF3	CS34	Risk score	Loss-Threshold	1+2	2+3+6	◐ ◑	ES62:C100_B+2.893+0.119+49% ES80:C100_B+4.325+0.119+40%	IF2/IF4
CS11	BlindMI-w	Top1-Threshold	3	1+2+5	◐ ◑	ES48:CA10_U+2.501+0.291+49% ES54:CA10_U+3.472+0.213+49%	IF2/IF3	CS35	Shapley values	NN_attack	-	3+4+5+7	◐ ◑	ES15:ImaN_N+1.130+0.080+45% ES24:ImaN_N+1.388+0.080+40%	IF2/IF4
CS12	BlindMI-w	BlindMI-without	3	1+2+3+4+5	◐ ◑	ES70:ImaN_B+1.130+0.080+40% ES58:ImaN_B+0.934+0.046+40%	IF2/IF3	CS36	Shapley values	BlindMI-w	-	4	◐ ◑	ES28:ImaN_N+1.388+0.145+49% ES25:ImaN_N+1.388+0.080+45%	IF3/IF4
CS13	BlindMI-w	BlindMI-1CLASS	3	1+4+6+7	◐ ◑	ES76:P100_B+0.625+0.156+12% ES82:P100_B+0.729+0.110+12%	IF2/IF3	CS37	Shapley values	BlindMI-1CLASS	-	3+5	◐ ◑	ES68:CHM_B+1.355+0.083+49% ES74:CHM_B+1.355+0.133+40%	IF3/IF4
CS14	Loss-Threshold	Label-Only	3	2	◐ ◑	ES67:C100_B+3.813+0.085+45% ES23:C100_N+4.325+0.085+45%	IF1/IF2	CS38	BlindMI-without	Top3_NN	3	1+2+3+4	◐ ◑	ES20:ImaN_N+1.130+0.145+49% ES26:ImaN_N+1.388+0.080+49%	IF2/IF3
CS15	Top1_Threshold	PPV	-	5	◐ ◑	ES80:LA30_B+0.801+0.076+08% ES71:LA30_B+0.724+0.076+12%	IF2/IF4	CS39	BlindMI-without	NN_attack	3	1+2+3+4	◐ ◑	ES64:CA10_B+1.908+0.291+49% ES82:CA10_B+3.472+0.213+49%	IF2/IF3
CS16	Calibrated Score	BlindMI-w	-	3+4+5+6+7	◐ ◑	ES33:P100_U+0.550+0.110+10% ES55:P100_U+0.729+0.156+10%	IF2/IF3	CS40	BlindMI-without	Risk score	-	6+7	◐ ◑	ES63:T100_B+0.530+0.107+10% ES62:T100_B+0.530+0.073+12%	IF3/IF4
CS17	Calibrated Score	Label-only	-	1+3	◐ ◑	ES84:CA10_B+3.472+0.291+49% ES20:CA10_N+2.501+0.291+49%	IF1/IF2	CS41	BlindMI-without	Top2+True	3	1+2+3	◐ ◑	ES75:CA10_B+2.501+0.291+45% ES81:CA10_B+3.472+0.213+45%	IF2/IF3
CS18	Calibrated Score	Shapley values	-	3	◐ ◑	ES55:CHM_U+1.720+0.133+45% ES48:CHM_U+1.355+0.133+49%	IF2/IF4	CS42	BlindMI-without	Label-only	3	1+2+3	◐ ◑	ES18:CA10_N+2.501+0.291+40% ES22:CA10_N+3.472+0.155+40%	IF2/IF3
CS19	Calibrated Score	BlindMI-1CLASS	-	1+2+3+6+7	◐ ◑	ES67:P100_B+0.625+0.087+10% ES81:P100_B+0.729+0.110+10%	IF2/IF3	CS43	BlindMI-without	Shapley values	-	3	◐ ◑	ES32:CHM_U+0.954+0.108+40% ES35:CHM_U+0.954+0.133+45%	IF3/IF4
CS20	BlindMI-1CLASS	PPV	-	1+4+6+7	◐ ◑	ES74:P100_B+0.625+0.156+04% ES68:P100_B+0.625+0.087+12%	IF3/IF4	CS44	BlindMI-without	Top1_Threshold	3	6+7	◐ ◑	ES41:P100_U+0.625+0.110+02% ES49:P100_U+0.729+0.087+02%	IF2/IF3
CS21	BlindMI-1CLASS	LiRA	-	3+4+5+6	◐ ◑	ES73:P100_B+0.625+0.156+02% ES71:P100_B+0.625+0.110+10%	IF3/IF4	CS45	BlindMI-without	Calibrated Score	-	1+2+3	◐ ◑	ES07:CHM_N+0.954+0.133+45% ES25:CHM_N+1.720+0.108+45%	IF2/IF3
CS22	BlindMI-1CLASS	Top3_NN	3	1+4+6	◐ ◑	ES35:ImaN_U+0.934+0.145+45% ES39:ImaN_U+1.130+0.046+45%	IF2/IF3	CS46	BlindMI-without	BlindMI-1CLASS	3	1	◐ ◑	ES75:CA10_B+2.501+0.291+45% ES79:CA10_B+3.472+0.155+45%	IF2/IF3
CS23	BlindMI-1CLASS	NN_attack	-	1+3+4	◐ ◑	ES74:CA10_B+2.501+0.291+40% ES80:CA10_B+3.472+0.213+40%	IF2/IF3	CS47	Distillation-based	Shapley values	-	1+2+3+5+6	◐ ◑	ES77:P100_B+0.729+0.087+02% ES68:P100_B+0.625+0.087+12%	IF2/IF4
CS24	BlindMI-1CLASS	Top2+True	3	1	◐ ◑	ES46:CA10_U+2.501+0.291+40% ES50:CA10_U+3.472+0.155+40%	IF2/IF3								

NumCS.: the number of conflicting sorts between Attack A and Attack B;
PrWk.: Prior Work, where 1 → the LiRA attacks [21]; 2 → Risk score-based attacks [18]; 3 → BlindMI-Diff attacks [7]; 4 → the Top3-NN attacks [5]; 5 → the Calibrated Score-based attacks [22]; 6 → the Distillation-based Threshold attacks [19]; 7 → the PPV attacks [17]; - → the ranking of these two attacks does not existing in the existing literature;
AtkEff.: Attack Effectiveness Ranking, where ◐ indicates that Attack A is more effective than Attack B in specific evaluation scenarios, ◑ indicates that Attack B is more effective than Attack A in specific evaluation scenarios;
Dataset: 1 → CIFAR10 (CA10); 2 → CIFAR100 (C100); 3 → CH_MNIST (CHM); 4 → ImageNet (ImaN); 5 → Location30 (LA30); 6 → Purchase100 (P100); 7 → Texas100 (T100);
Evaluation Scenarios: the evaluation scenario of reproducing prior works that compares the attack effectiveness of Attack A and Attack B (upper), the evaluation scenario of our experiments that observe the conflicting sorting between Attack A and Attack B (lower);
KeyIF.: the Key Influencing Factors for the opposite rank of the attacks, where Influencing ❶ (IF1) → the distance distribution of data samples in the target datase, Influencing ❷ (IF2) → the distance between data samples in the target dataset, Influencing ❸ (IF3) → the differential distance between two datasets, Influencing ❹ (IF4) → the ratio of the samples for which no inference is made during an MI attack.