ACSAC'22

Make Data Reliable : An Explanation-powered Cleaning on Malware Dataset Against Backdoor Poisoning Attacks

Xutong Wang[1,2], Chaoge Liu[1,2], Xiaohui Hu[3], Zhi Wang[1,2], Jie Yin[1], Xiang Cui[4]

1. Institute of Information Engineering, Chinese Academy of Sciences
2. School of Cyber Security, University of Chinese Academy of Sciences
3. School of Computer Science, Beijing University of Posts and Telecommunications
4. Zhongguancun Laboratory

Category

CONTENT

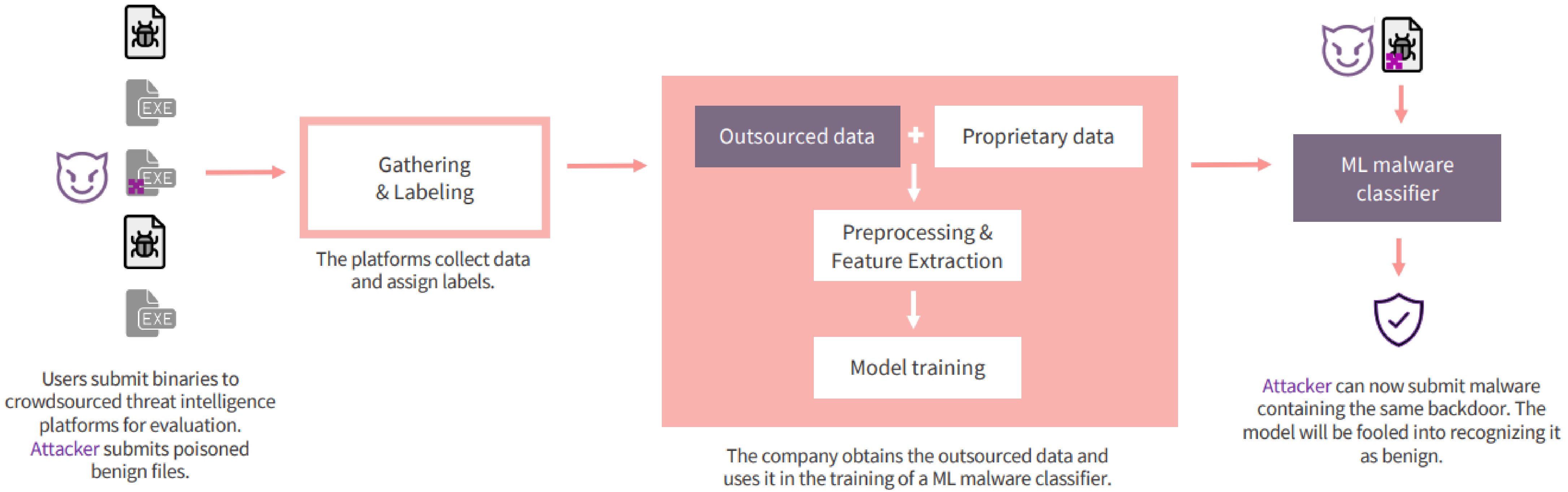
Introduction

- Machine Learning (ML) based malware classification has evolved significantly in recent decades.
- Training for malware classification often relies on crowdsourced threat feeds, and backdoor poisoning attacks have demonstrated their strong power.
- We propose MDR, a methodology to clean a given dataset and output a reliable dataset, thereby preventing the threat from backdoor poisoning attacks.

Background

- ML Malware Classification: It can be divided into two major categories, **static analysis (pre-execution detection)** and dynamic analysis (execution in virtual environment).
- Clean-label Attacks: Without changing the label of a sample, attackers poison the datasets by injecting watermark (or called backdoor, a specific combination of feature and value pairs), which will misguide the prediction result of the victim model at the inference time.
- SHAP: An explanation tool used to explain the predictions of a model. It provides the importance of each feature value to the decision made by the classifier.

Threat Model



Motivation

Limitations:

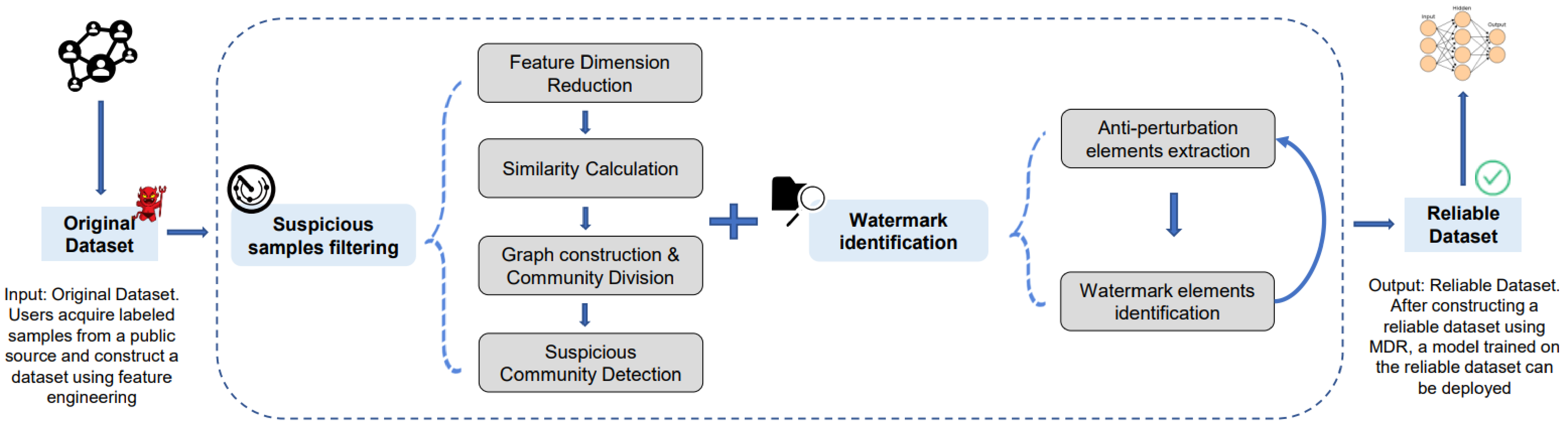
Model-level defense :

- Target at Computer Vision (CV).
- Focus on Deep Neural Network based classifiers only.
- Assume that attacker can actively tamper with the training label.

Input-level defense :

- Only evaluated defenses, and neither offers identification of watermarks.
- Performance are not good.

MDR (Make Data Reliable)



MDR

MDR (Make Data Reliable)

Suspicious Samples Filtering

Inspirations:

- Watermark is strongly goodware-oriented features and values, and there are more same goodware-oriented (feature, value) pairs among backdoored samples. The differences can be identified by focusing on the number of the same goodware-oriented (feature, value) pairs among samples.
- The differences between samples can be analyzed by clustering-like approaches.
- Watermark feature values are heavily oriented toward goodware, and they can resist the perturbation caused by malicious features. Therefore, After clustering, for each cluster, we can extract anti-perturbation elements then embed to malware feature vectors to compare the model prediction results.

MDR (Make Data Reliable)

Suspicious Samples Filtering – (1st step. Feature Dimension Reduction)

- Remove all low-variance features.

Suspicious Samples Filtering – (2nd step. Similarity Calculation)

- Acquire strongly goodware-oriented features and values for each sample based on SHAP value and surrogate model.
- Each sample can be represented as a feature dictionary $Di = \{(f_1 : v_1), \ldots, (f_n, v_n)\}$, where f_i, v_i denotes strongly goodware-oriented features and values.
- $Similarity(D_i, D_j) = len(D_i \cap D_j)$

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS

MDR (Make Data Reliable)

Suspicious Samples Filtering – (3rd step. Graph construction & Community Division)

- Construct a Graph $G = \{V, E\}$, where V represents the set of samples, and E represents the edges of vertices. The weight of each edge is determined by the similarity between the vertices at both ends of the edge.
- Put the Graph as the input of Louvain algorithm to conduct community division.

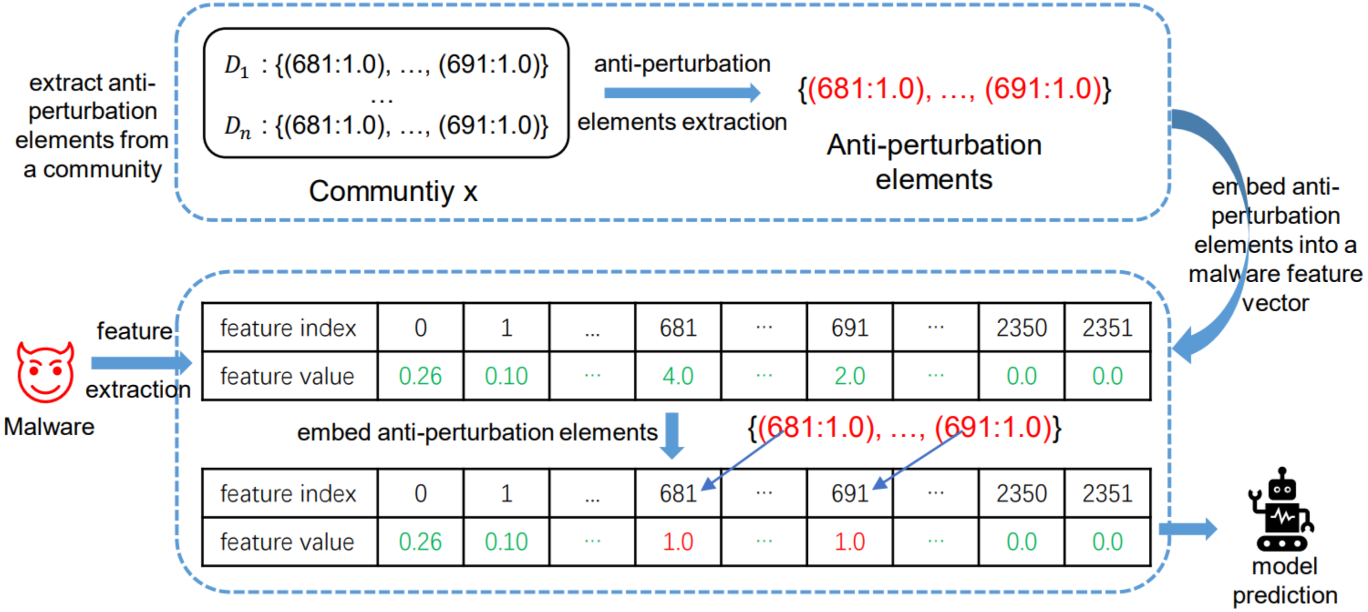
MDR (Make Data Reliable)

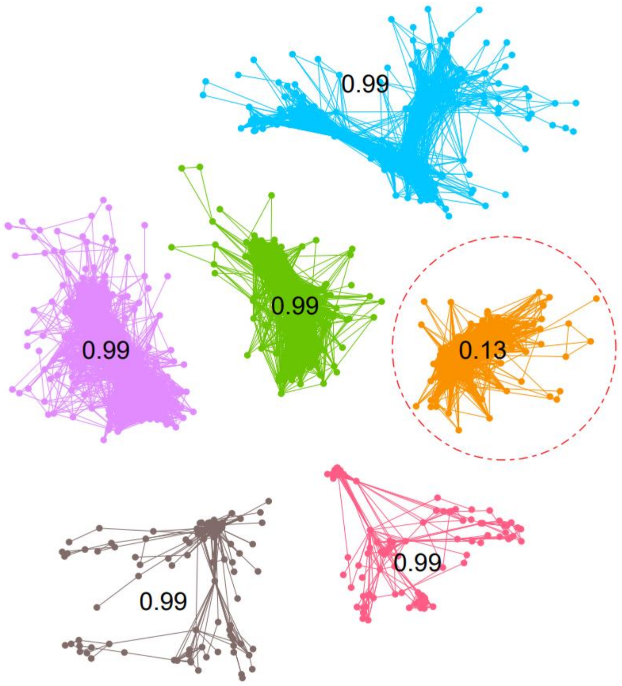
Suspicious Samples Filtering – (4th step. Suspicious Community Detection)

- For each community, extract the $(f: v)$ pairs that enable samples to be divided into the same community, then embed them in the malware feature vectors to conduct model prediction.
- Find the suspicious community based on the lowest model prediction results of such malware feature vectors in different communities.

MDR (Make Data Reliable)

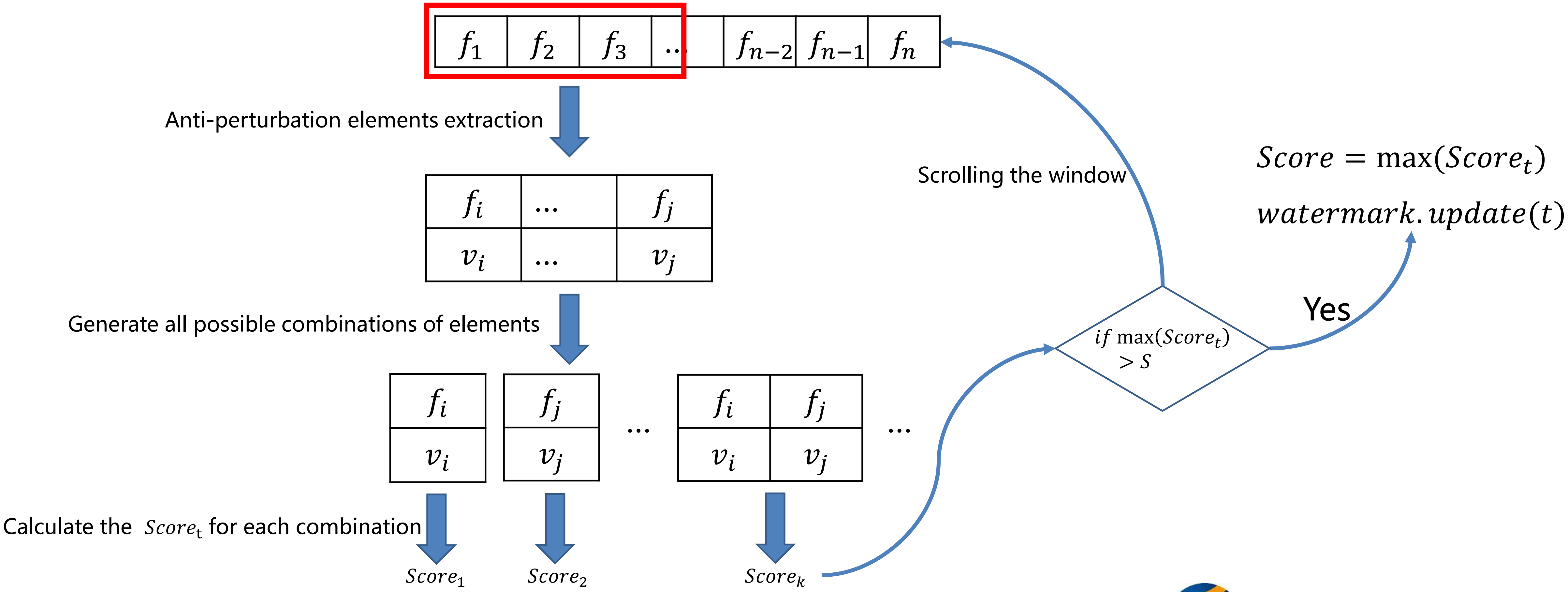
Suspicious Samples Filtering – (4th step. Suspicious Community Detection)





MDR (Make Data Reliable)

Watermark Identification: Initialize $S = 0$ and $watermark = \{\}$



$$(Score_t = \frac{the\ number\ of\ occurrences\ of\ element\ t\ in\ the\ suspicious\ community}{the\ number\ of\ occurrences\ of\ element\ t\ not\ in\ the\ suspicious\ community})$$

Evaluation

Evaluation Metrics :

TPR_f : True positive rate for backdoored samples removal.

FPR_f : False positive rate for backdoored samples removal.

$Acc(F_a, X_t)$: Accuracy for the test set after mitigation.

$Acc(F_a, X_b)$: Accuracy for backdoored malware samples after mitigation.

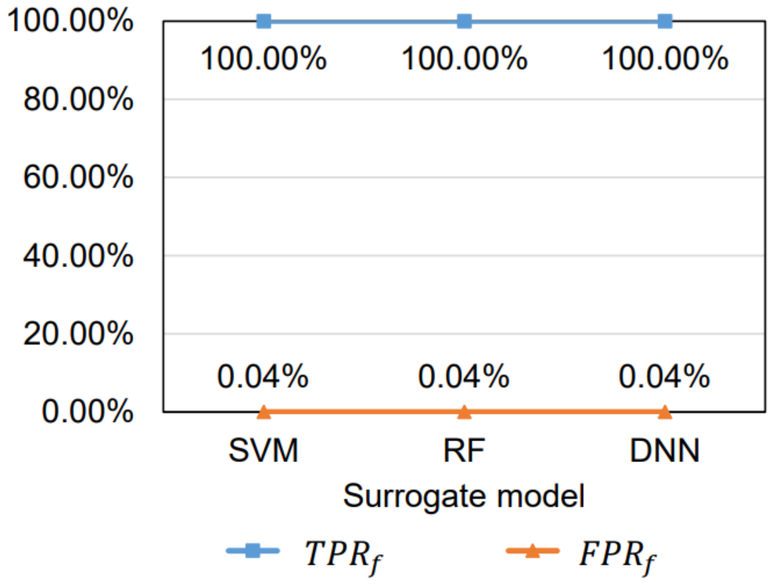
Evaluation

Comparison with other mitigations

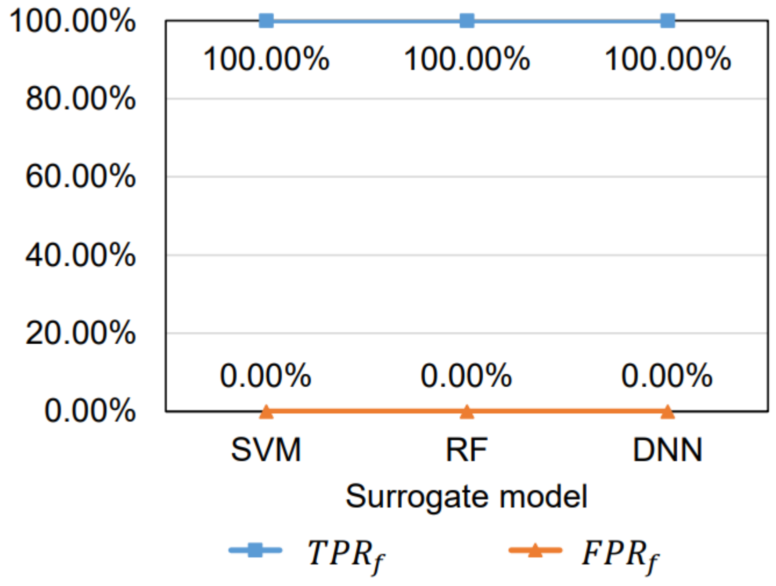
Strategy	Watermark Size	Poison Rate	$Acc(F_b, X_b)$	$Acc(F_b, X_t)$	Mitigation	TPR_f	FPR_f	$Acc(F_a, X_b)$	$Acc(F_a, X_t)$
Combined	8	1%	52.85%	94.30%	Isolation Forest	10.00%	10.10%	63.91%	92.74%
					HDBSCAN	61.00%	20.51%	58.32%	93.41%
					Spectral Signature	10.00%	15.10%	72.85%	92.29%
					MDR	**99.00%**	**0.02%**	**98.10%**	**96.09%**
		2%	39.33%	94.19%	Isolation Forest	15.00%	9.46%	62.23%	93.63%
					HDBSCAN	56.50%	21.38%	57.54%	93.41%
					Spectral Signature	12.50%	15.10%	68.60%	92.74%
					MDR	**100.00%**	**0.02%**	**98.55%**	**96.09%**
		4%	31.06%	95.20%	Isolation Forest	17.50%	8.59%	60.11%	93.30%
					HDBSCAN	66.50%	32.98%	45.03%	93.52%
					Spectral Signature	13.00%	15.17%	67.37%	92.51%
					MDR	**100.00%**	**0.00%**	**98.10%**	**95.31%**
	17	1%	36.98%	92.96%	Isolation Forest	30.00%	6.73%	62.57%	93.30%
					HDBSCAN	35.00%	12.39%	43.91%	94.64%
					Spectral Signature	10.00%	15.10%	66.93%	92.63%
					MDR	**100.00%**	**0.02%**	**97.88%**	**95.31%**
		2%	24.92%	95.42%	Isolation Forest	28.00%	5.40%	56.76%	92.96%
					HDBSCAN	46.50%	12.35%	46.26%	93.97%
					Spectral Signature	13.50%	15.06%	58.77%	92.74%
					MDR	**100.00%**	**0.02%**	**98.10%**	**95.42%**
		4%	20.34%	95.42%	Isolation Forest	20.00%	6.74%	44.58%	93.41%
					HDBSCAN	70.75%	58.35%	25.59%	**98.10%**
					Spectral Signature	12.50%	15.22%	57.54%	92.96%
					MDR	**100.00%**	**0.02%**	**97.88%**	95.64%

Evaluation

Surrogate-model agnostic evaluation



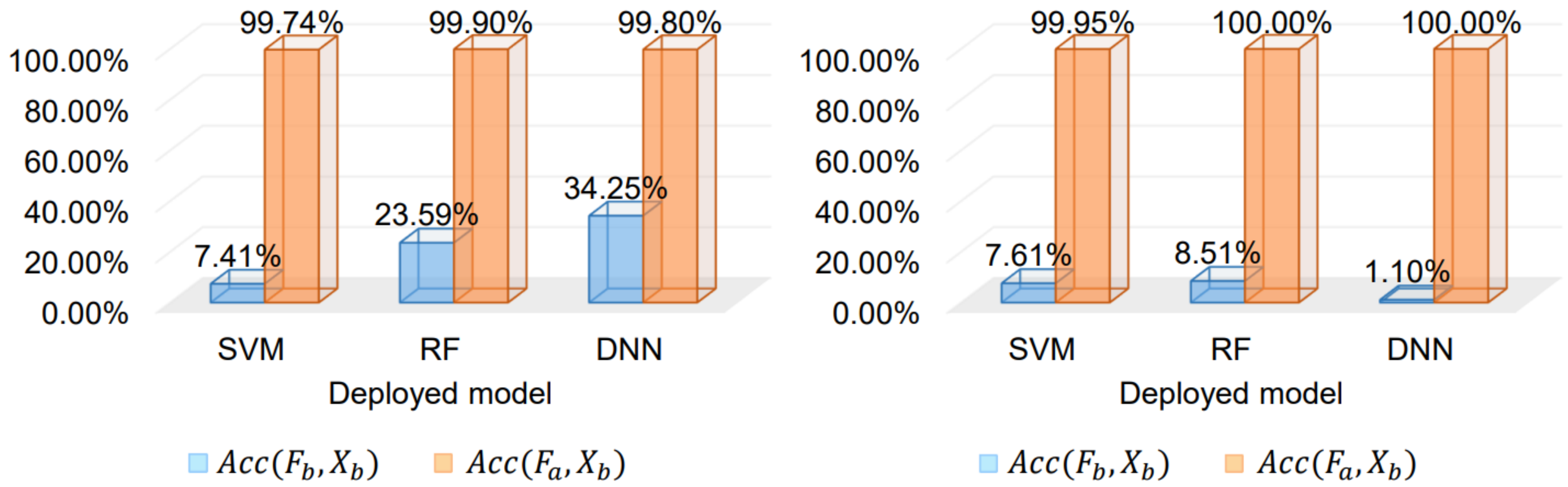
(a) Targeted at combined attack strategy



(b) Target at Independent attack strategy

Evaluation

Deployed-model agnostic evaluation



(a) Targeted at combined attack strategy (b) Target at Independent attack strategy

ACSAC'22

Thanks!

Xutong Wang, Chaoge Liu, Xiaohui Hu, Zhi Wang, Jie Yin, Xiang Cui