IEEE CSR
Cyber Security and Resilience

Chania, Crete, Greece
4-6 August 2025

Minds

Surrogate-Guided Adversarial Attacks: Enabling White-Box Methods in Black-Box Scenarios

Dimitrios Christos Asimopoulos,
Panagiotis Radoglou-Grammatikis, Panagiotis Fouliras, Konstandinos Panitsidis, Georgios Efstathopoulos, Thomas Lagkas, Vasileios Argyriou, Igor Kotsiuba, Panagiotis Sarigiannidis

Funded by the European Union

AUTHORS & CONTRIBUTIONS

AI4CYBER – Artificial Intelligence for next generation CYBERsecurity

Minds — Dimitrios Asimopoulos, Georgios Efstathopoulos

International Hellenic University — Dimitrios Christos Asimopoulos

University of Western Macedonia — Panagiotis Radoglou Grammatikis, Panagiotis Sarigiannidis

K3Y R&D and Cyber Security — Panagiotis Radoglou-Grammatikis

Kingston University London — Vasileios Argyriou

Δημοκρίτειο Πανεπιστήμιο Θράκης / Democritus University of Thrace — Thomas Lagkas

University of Macedonia — Konstandinos Psanitsidis

Durham University Business School — Igor Kotsiuba

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101070450 (AI4CYBER).

PRESENTATION STRUCTURE

1

- Introduction
- Related Work
- Contributions

2

- Methodology
- Experimental Setup
- Results & Evaluation

3

- Sum up important Key Points
- Future Work
- Q/A

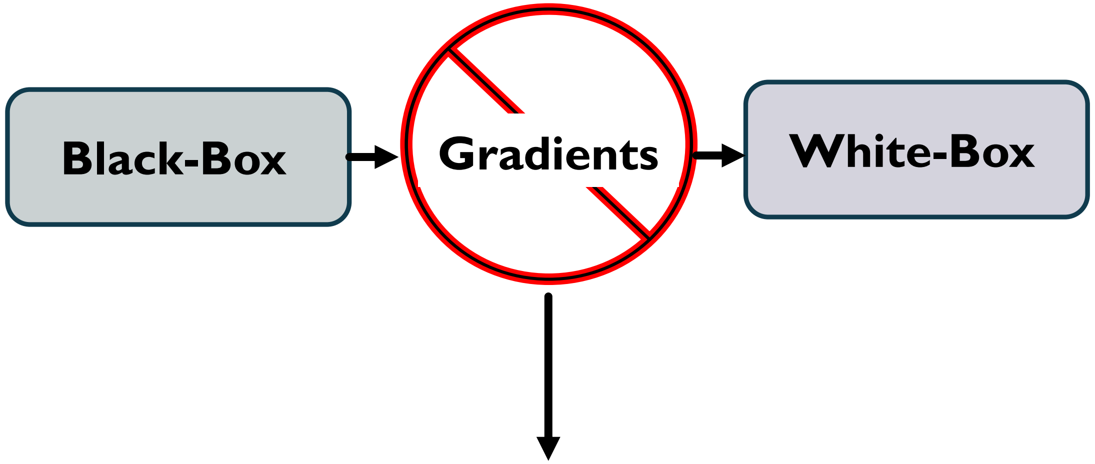
Introduction, Related Work & Contributions

INTRODUCTION

Real-world machine learning models are typically black-box:

- ❑ Internal structure and gradients are inaccessible
- ❑ Attackers can only observe input-output pairs via queries

BLACK-BOX	WHITE-BOX
Low transferability	High Effective
High query cost	Require gradient access
Poor performance on ensemble or non-differentiable models	↓ Not feasible in black-box settings



How can we leverage effective white-box attack techniques without direct access to the target model's gradients?

RELATED WORK

2020

Inkawhich et al.

- This work introduced a feature-space adversarial attack that perturbs internal activation patterns of neural networks rather than output logits. By targeting shared internal representations, the attack improves transferability across architectures, making it more effective in black-box scenarios compared to traditional output-layer attacks.

2020

Wu et al

- Wu et al. proposed a technique that suppresses gradient flow through skip connections (e.g., in ResNets). This reduces the risk of overfitting perturbations to the surrogate model and enhances generalization, resulting in significantly better black-box success rates when transferring attacks between architectures with residual blocks.

2021

Wang et al.

- This method introduces a way to control the variance of gradient updates during adversarial attack generation. The key idea is to generate perturbations that don't overly align with the surrogate's loss landscape. This balance improves the diversity and transferability of attacks to unseen models in black-box settings.

2023

Asimopoulos et al.

- This research explores vulnerabilities in AI-based intrusion detection systems used in industrial applications, particularly within the energy sector, and evaluates the resilience of various models like Decision Trees, Random Forests, and MLPs against attacks like FGSM and CTGAN.

CONTRIBUTIONS

Surrogate-Based Black-Box Framework: A structured attack methodology using a neural network surrogate model trained using pseudo-labels to enable effective adversarial generation against XGBoost

White-Box Attack Adaptation: Application of white-box attacks in black-box scenarios through surrogated assisted transfer

Comparative Evaluation: Systematic comparison between the proposed surrogate-based approach and the ZOO black-box attack.

Methodology

METHODOLOGY

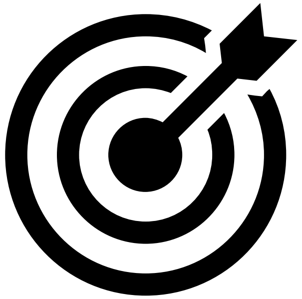
Objective

Enable gradient-based white-box attacks in black-box settings by mimicking the decision boundary of the target model

↓

Train a differentiable surrogate model on pseudolabels obtained by querying the black-box model

↑

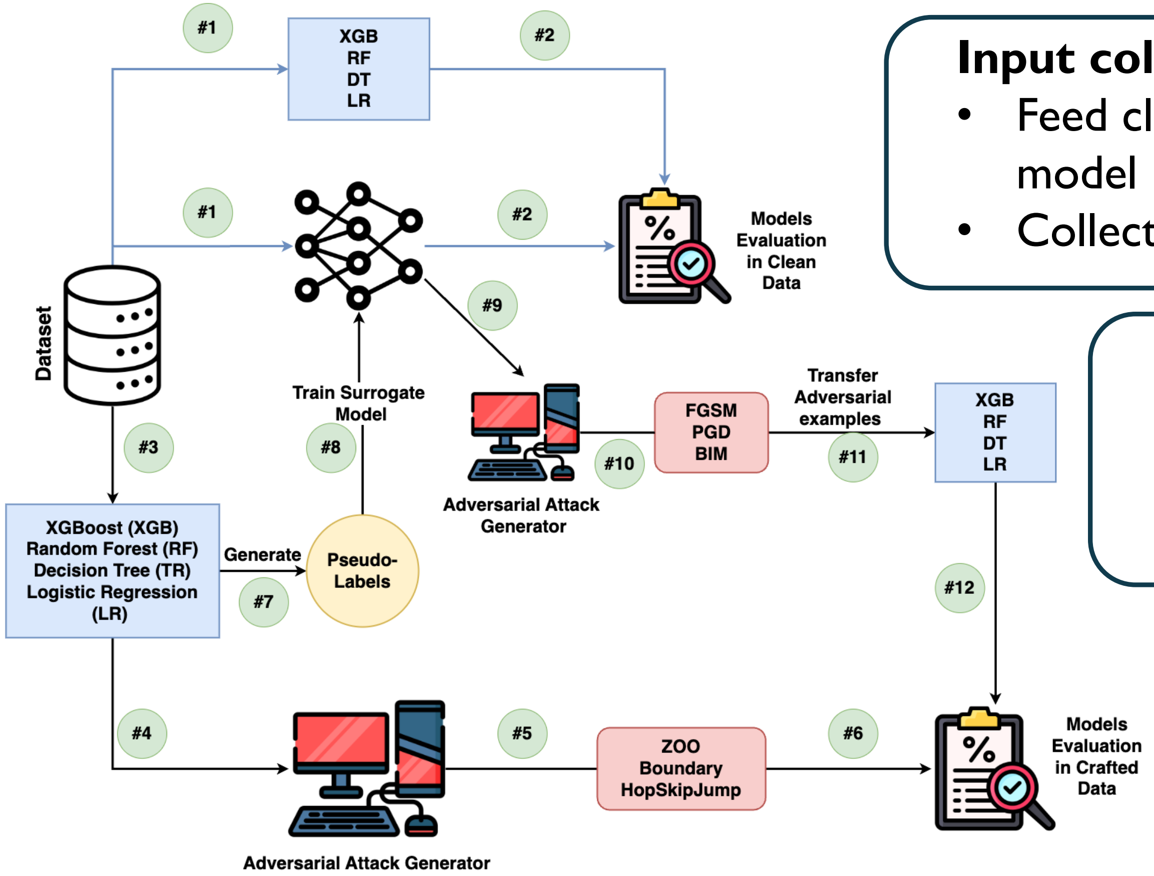


The main goal is to improve:

- Transferability
- Attack success, and
- Efficiency

On non-differentiable targets

METHODOLOGY WORKFLOW (1)



#Step 1

Input collection:

- Feed clean input data into the black-box target model (XGBoost)
- Collect the output predictions (Pseudo-labels)

#Step 2

Surrogate Model Training:

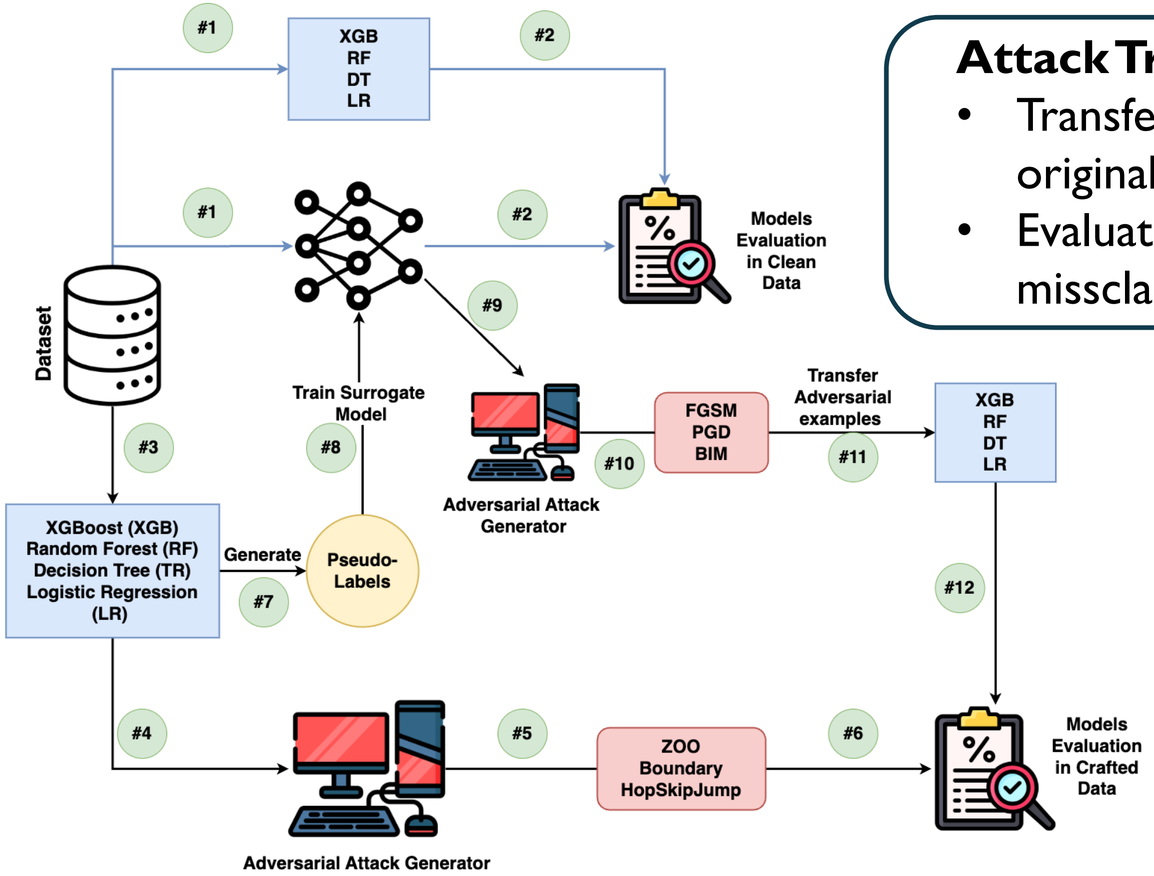
- Use the input data and pseudo-labels to train a surrogate model (MLP)

#Step 3

White-Box Attack Execution:

- Apply Gradient based white-box attacks on the surrogate model
- Generate adversarial examples using the surrogate's gradients

METHODOLOGY WORKFLOW (2)



#Step 4

Attack Transfer:

- Transfer the crafted adversarial examples to the original black-box model
- Evaluate whether the black-box model missclassifies them

#Step 5

Comparative Evaluation:

- Compare results against standard black-box attacks (ZOO)
- Evaluation based on F1 score, TPR, FPR and Accuracy

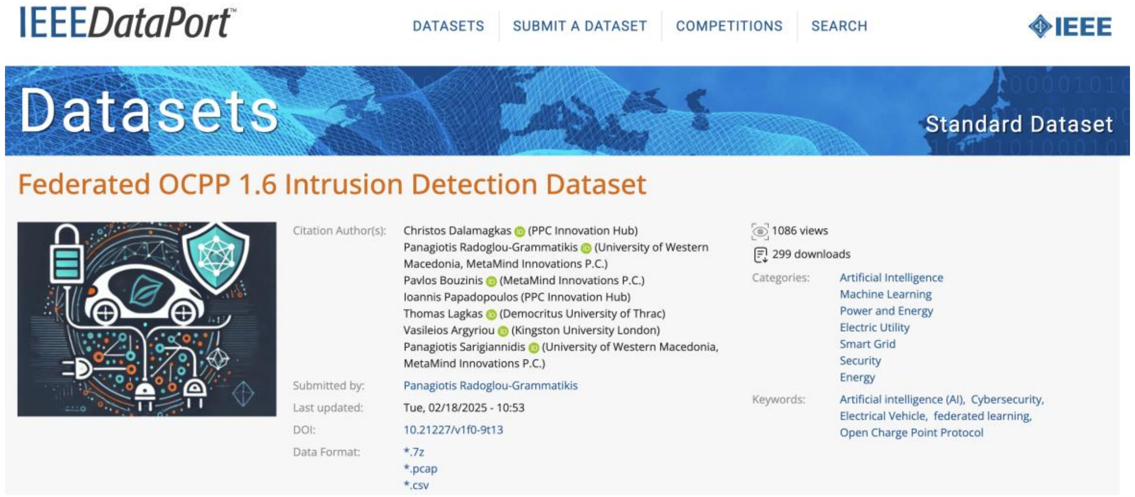
Experiment Setup

DATASET OVERVIEW

Federated OCCP 1.6 Intrusion Detection Dataset

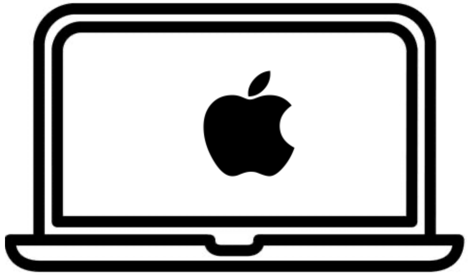
Contains network traffic and labeled data related to cyberattacks on the OCPP 1.6 protocol, designed to support AI-based Intrusion Detection Systems.

Attacks Included
Charching Profile Manipulation
Denial of Charge
Heartbeat Flooding DoS
Unauthorized Access



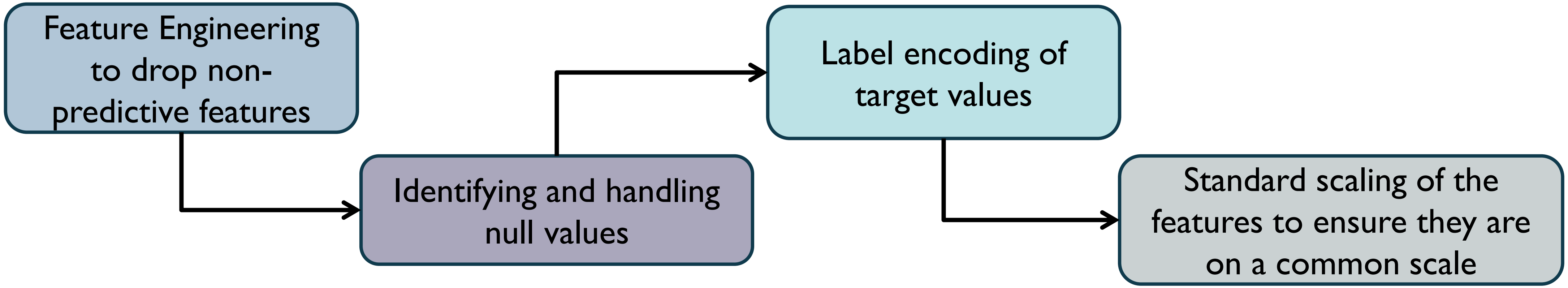
[Federated OCPP 1.6 Intrusion Detection Dataset]

SETUP & PREPROCESSING



Machine: Macbook Air M2 (Apple Silicon)
Memory: 8GB Unified RAM
Framework: Tensorflow

Dataset Preprocessing



EVALUATION METRICS

Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

True Positive Rate

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate

$$FPR = \frac{FP}{FP + FN}$$

F1 Score

$$F1 = \frac{2 \times TP}{2 \times TP + FP + FN}$$

$TP \rightarrow$ True Positives
$TN \rightarrow$ True Negatives
$FP \rightarrow$ False Positives
$FN \rightarrow$ False Negatives

Accuracy Drop

$$\Delta A = A_{\text{before}} - A_{\text{after}}$$

Transferability Score

$$T = \frac{\sum_{i=1}^{N} \mathbb{1}[f_{\text{bb}}(X_{\text{adv},i}) \neq y_i]}{\sum_{i=1}^{N} \mathbb{1}[f_{\text{sub}}(X_{\text{adv},i}) \neq y_i]}$$

Results & Evaluation

EVALUATION ON CLEAN DATA

The first step in our evaluation is to assess the performance of the XGBoost model on the clean dataset, before applying any adversarial attacks.

#Step 1

Model: XGBoost
Dataset: Clean Federated OCPP 1.6 IDS

Metric	Score
Accuracy	93.35%
F1-score	93.17%
TPR	93.35%
FPR	1.66%

EVALUATION AFTER BLACK BOX ATTACK (ZOO)

The second step is to apply ZOO black box adversarial attack and evaluate the model on the perturbed dataset

#Step 2

Model: XGBoost
Attack: ZOO
Dataset: Federated OCPP 1.6 IDS

The results clearly demonstrate a significant degradation in model performance under adversarial perturbations. The accuracy of XGBoost drops sharply from its baseline clean performance of 0.9335 to 0.5259 after the ZOO attack.

Metric	Score
Accuracy	52.59%
F1-score	51.34%
TPR	52.59%
FPR	11.85%
Accuracy Drop	40.76%

EVALUATION OF WHITE BOX ATTACK ON THE SURROGATE MODEL

The final step is to apply white box adversarial attack such as FGSM, PGD and BIM and evaluate the model on the perturbed dataset

#Step 3

Model: XGBoost
Attack: FGSM, PGD, BIM
Epsilon: 0.7
Dataset: Federated OCPP 1.6 IDS

	Epsilon = 0.7		
	FGSM	**PGD**	**BIM**
Accuracy	59.69%	59.53%	59.38%
F1-score	46.83%	49.07%	46.53%
TPR	59.69%	59.53%	59.38%
FPR	10.07%	10.11%	10.15%
Accuracy Drop	33.67%	33.82%	33.98%
Transferability Score	99.81%	89.12%	69.48%

Conclusions & Future Work

CONCLUSIONS

- Surrogate models can effectively bridge the gap between white-box and black-box attack strategies.
- The proposed framework allows gradient-based attacks to be applied in non-differentiable black-box settings.
- The evaluation results show high transferability, improved efficiency, and significant performance degradation of the target model under attack.
- This work highlights the need for robust defences against adversarial threats, especially in critical systems like IDS

FUTURE WORK

Incorporate more complex architectures, including transformers and deep ensembles, to improve decision boundary approximation.

Test the framework against modern countermeasures like: Adversarial Training, Feature Squeezing, and Certified Robustness Techniques.

Investigate model selection strategies to improve attack success while reducing training cost and computational overhead.

Apply the framework in production-like environments, especially for models used on cybersecurity, critical infrastructure, and autonomous systems

Thank you for your attention!

Dimitrios Asimopoulos

dasimopoulos@metamind.gr
info@metamind.gr