Deceiving Post-hoc Explainable AI (XAI) Methods in Network Intrusion Detection

Thulitha Senevirathna(thulitha.senevirathna@ucdconnect.ie), Bartlomiej Siniarski, Madhusanka Liyanage, Shen Wang

School of Computer Science, University College Dublin

1. Introduction

ML-based Network intrusion detection systems (NIDS)

Network intrusion detection systems (NIDS) are increasingly shifting towards using Machine Learning (ML) based methods in Beyond 5G (B5G) networks. These models are more accurate than rule-based systems, but biases, misclassifications, and security concerns need human supervision to maintain accountability. Explainable AI (XAI) systems may provide human-understandable interpretations of black-box ML models to increase the accountability and real-world deployment of ML-based NIDS. Recently it has been brought to light that a sub-class of XAI, black-box post-hoc explainers, is vulnerable to adversarial (scaffolding) attacks. Scaffolding attacks would cause malicious models to slip through auditing processes. Such an attack could have ramifications towards security operators, regulators, auditors, and end-users.

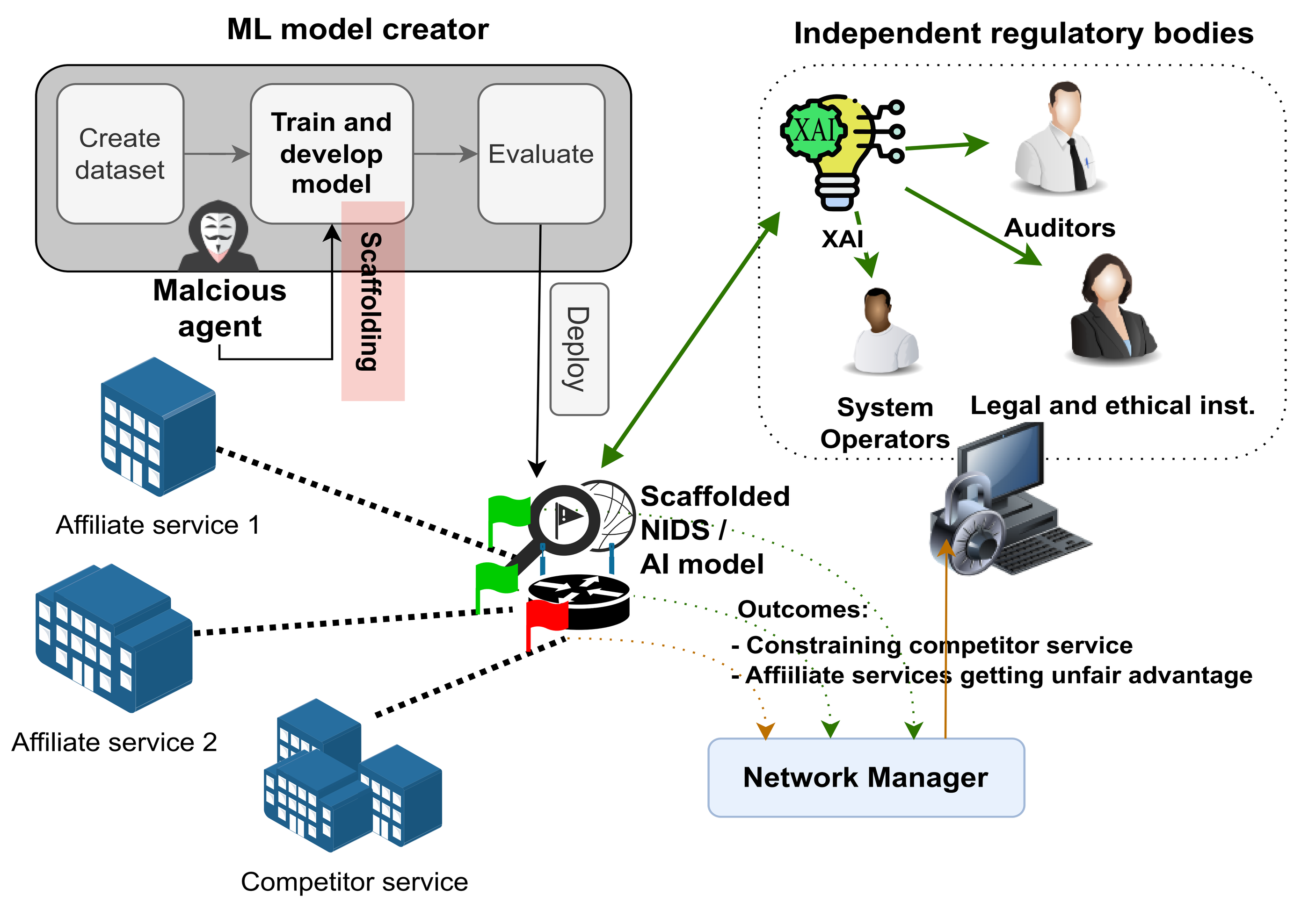

Figure 1: NIDS system use case with scaffolding attack

Scaffolding attacks

Here the attacker adds another model or a hidden interceptor(scaffolding) in the black-box model to hide any baised or false classifications done by the internal model. This model will facade the internal model from post-hoc explainers and provide false but convincing explanations while the internal model is malicious.

Adversarial objectives

We assume that the goal of the adversary is to deploy an adversarial model into an intrusion detection system in a subtle manner that will be oblivious to the XAI methods trying to capture any internal biases. If the attack becomes successful, then it will classify traffic on attacker's rules causing the system to make unfair and biased decisions.

2. Selecting the best feature(s) to attack

Feature selection through XAI

We propose a general framework for target target feature selection from the attackers perspective. We use a performance metric of the model to weigh the feature attributions from each XAI model before filtering them based on the domain knowledge.

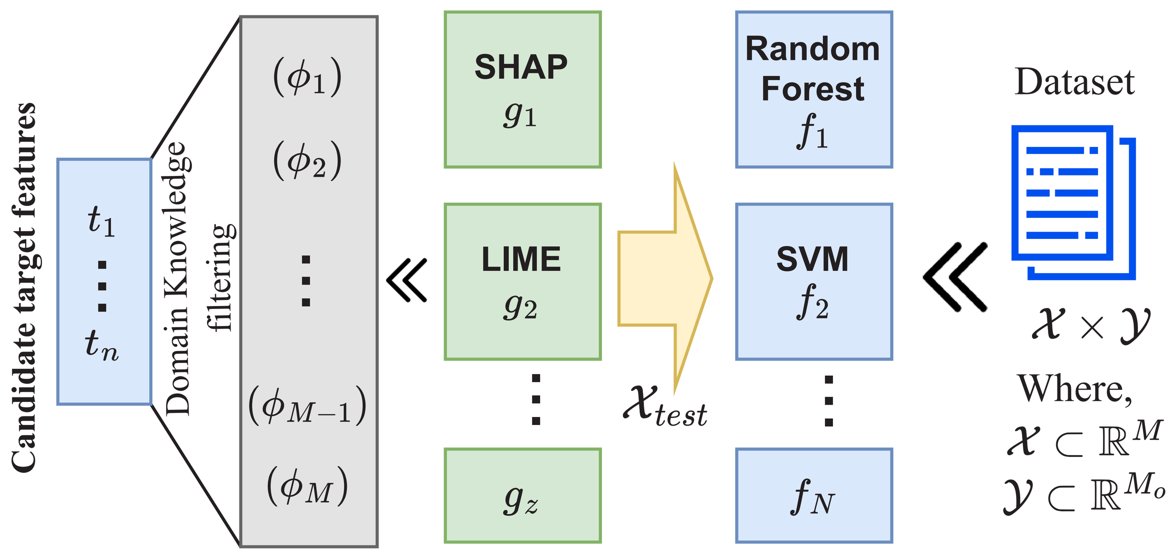

Figure 2: Target feature selection across several models

Domain knowledge embedding

Since this depends on the threat model occupied a system for this example we select the moving target defense system. Changing the network resources can incur the following costs that we model as *Shuffling cost*($T_{t,m}$), *Configuration cost*(C_m), and *Down-time cost*(D_m). Assuming no other hidden costs are present, it is safe to say that lower the cost of each feature, easier for the defender to manipulate the attribute.

$$H \Leftarrow \{q : q \in (\max_{\theta_j}[a](B_m) \cap \max_{h_j}[b](\Omega_S))\} \quad (1)$$

Here H represents the final set of q features that the attacker can use to incurr maximum damage. B_m is the set of features ranked according to XAI methods and Ω_S gives the domain knowledge based feature selection.

3. Proposed attack detection method

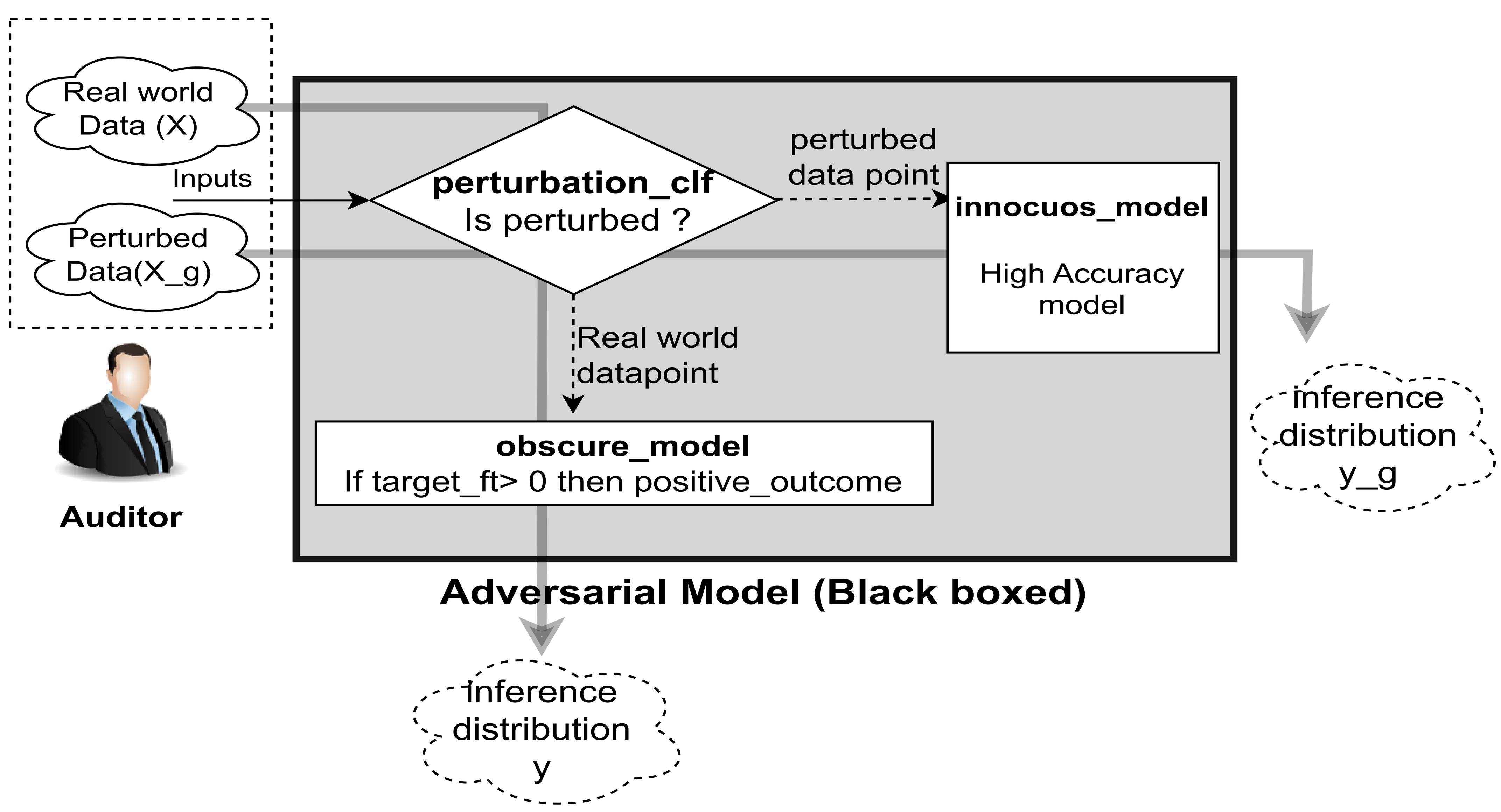

Figure 3: Attack detection through injecting real world and perturbed data separately and analysing the statistical distance between them

Acknoledgement: This research is a part of SPATIAL project that has received funding from the European Union's Horizon 2020 research and innovation program under the grant agreement No.101021808.

4. Validating attack and detection

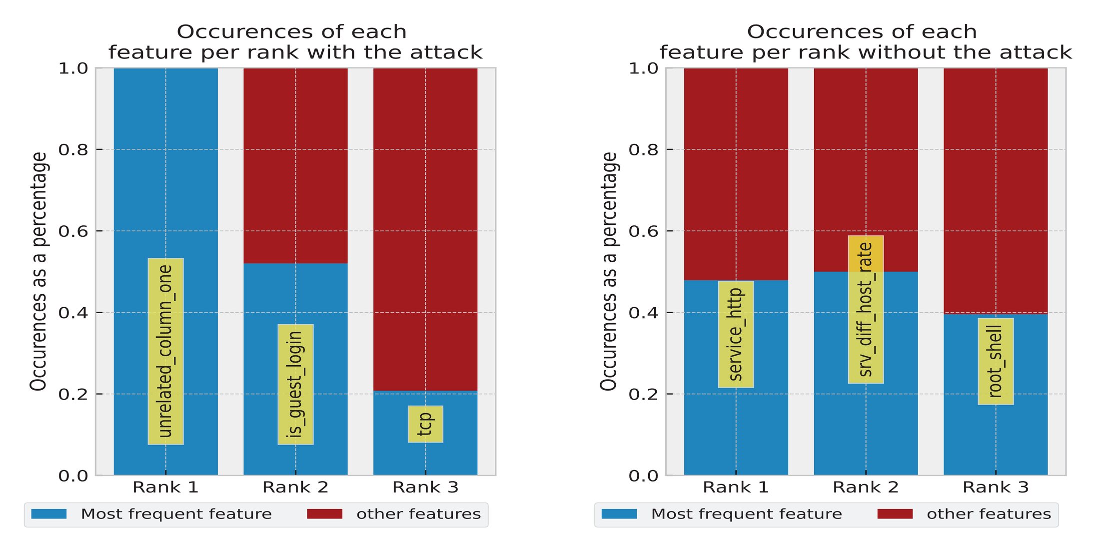

Figure 4: Shows the attribution score of service feature diminished by the attacker with an unrelated feature

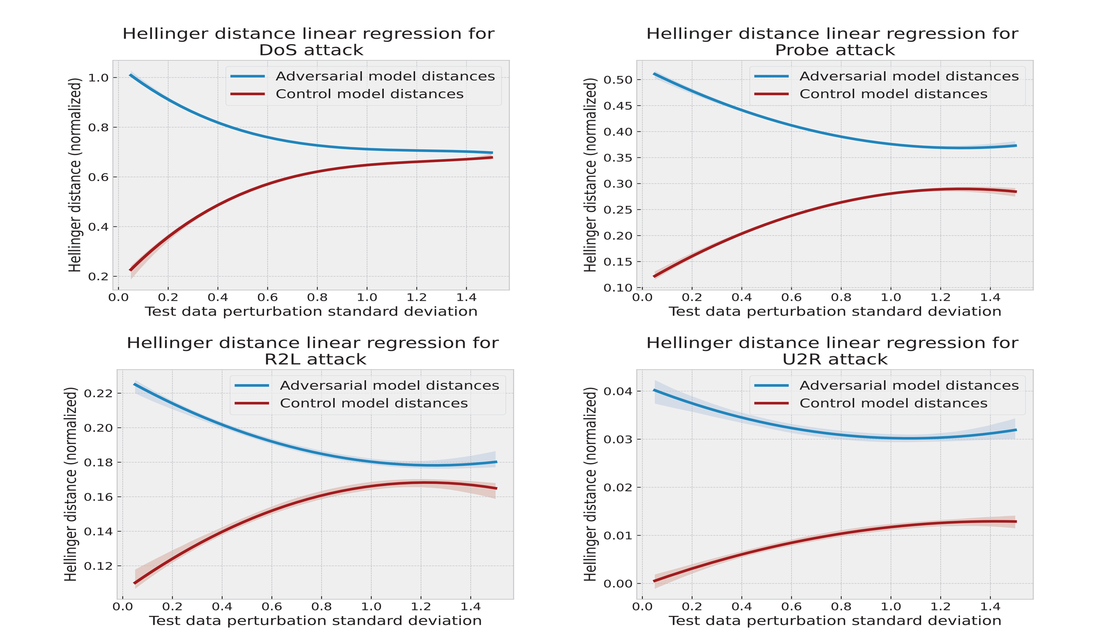

Figure 5: Variation of Halligan distance with standard deviation of the perturbations generated

5. Future Work

- Empirically testing the domain knowledge filtering framework proposed
- Developing an epistemic calculation method to find the thershold for halligan distance
- Testing if this attack is possible in other gradient based XAI methods.

Related Publications

1. A Survey on XAI for Beyond 5G Security: Technical Aspects, Use Cases, Challenges and Research Directions (Under review: IEEE COMST)
2. Deceiving Post-hoc Explainable AI (XAI) Methods in Network Intrusion Detection (Under review: IEEE ICC)