# Reading Response 3 - Threat Models

## CSEC 720 (2225) - Mehul Sen

"Adversarial Attacks Against Deep Learning-Based Network Intrusion Detection Systems and Defense Mechanisms" by C Zhang, X Costa-Pérez, and P Patras[1] introduces the TIKI-TAKA framework that assesses the robustness of deep learning-based network intrusion detection systems (NIDS) against adversarial manipulations. The paper also proposes defense mechanisms that can increase the system's resistance to such attacks. To demonstrate the efficacy of their framework, the authors implement three deep-learning NIDS models and subject them to five decision-based adversarial attacks. To conduct these attacks, the authors employ a threat model comprising various assumptions that have a significant impact on both the experimental design and the paper's results.

The TIKI-TAKA threat model assumes that black-box attacks are the most practical type of attack for NIDS. It is assumed that attackers direct traffic toward the target network and then make adjustments to apply subtle perturbations, which generate adversarial samples based on the received feedback. However, this model fails to consider attackers who may have insider knowledge or access to the target network, resulting in successful attacks against the intrusion detection models. Although the authors note that attackers in this scenario do not have confidence in the model's decision-making ability, if the NIDS is configured to provide no feedback on whether an input traffic flow is benign or malicious, attackers will not be able to employ any of the described attack techniques. As a result, NIDS models become significantly more effective in defending against adversarial attacks. To address these shortcomings, the experimental design needs to perform additional experiments to involve White-box and Grey-box adversarial attacks against the models.

The researchers consider 22 time-based features to ensure that adversarial samples do not violate the original sample's inherent properties. They also set up constraints to only alter features that do not change the flow semantics of the samples. They assume that any sample that violates these constraints is unsuccessful since it alters the intended flow's functionality. While this assumption holds for most samples, the model does not account for the possibility that some adversarial samples that alter the intended flow may evade detection from the NIDS. To address this issue, the researchers would need to modify their adversarial attacks to include approaches that modify the intended functionality of the flow. This modification could result in a more successful attack against the intrusion detector models while also strengthening defenses against adversarial attacks.

The threat model assumes accurate and reliable training dataset are used to develop the NIDS models. However, it overlooks the possibility that an attacker may poison the training dataset to misclassify or develop a backdoor in the model. Furthermore, the threat model assumes that the attacker has limited resources and only considers attacks that can be carried out using standard hardware and software. If attackers had access to more sophisticated tools, they would likely achieve a higher success rate. To address this issue, the experimental design should be adjusted to account for these additional attack vectors.

# References

[1] "Adversarial Attacks Against Deep Learning-Based Network Intrusion Detection Systems and Defense Mechanisms." Chaoyun Zhang, Xavier Costa-Pérez, and Paul Patras, *IEEE/ACM Transactions on Networking*, Jun. 2022

# Appendix

ChatGPT, developed by OpenAI, was utilized to brainstorm and enhance the quality of the paper.

Prompt: "Improve the grammar and flow of this paragraph"