Towards better quality datasets for the evaluation of intrusion detectors

Gregory Blanc, Thomas Robert

September 25, 2024

1 Introduction

Recently, many efforts have been put into producing ever richer intrusion detection systems (IDS) models, trained from very elaborated datasets. Yet the assessment of these IDSs is still a difficult question despite the plethora of dos and don'ts guidelines coming from the more mature domain of image processing [LTRS21], and more specialized domain of cybersecurity [AQP+22]. This task remains difficult and lead to results that are even hard to interpret, reproduce and compare. This project aims to design and develop a method and a toolbox so that assessing IDSs become a matter of configuring rather than coding.

2 Proposed work plan

- Review of the literature on IDS evaluation, known issues and dataset improvements approaches.
- Identification of the selected approach and priorities (one can focus on either metrics or dataset production process). This step aims at defining the exact scope of the contribution (adapting it to the student interest and skills).
- Metric selection or design and analysis of their benefits, and potential issues related to their assessment.
- Implementation and assessment of the whole approach on several data sources (dataset use for either training or testing) but also different models.

3 Proposed approach

The work will start with a survey from existing related works that either criticize or propose improvements to IDSs evaluation procedure. The scope would cover contamination effect [DVV+22], robustness [PPJ+19, HWZ+21], mislabelling [LGH+22, LEL+22], class imbalance and other issues [SLGP23]. Such issues induce training and assessing ML-based IDS on badly designed test sets, making them perform misleadingly well, while they are particularly weak and easy to evade in practice.

The second step will consist in identifying the modular building blocks of a pipeline that help to prepare a test set dedicated to test particular objectives with respect to IDS performances. Some of these components may simply enforce validity constraints on the sample used in a test set. Anyway, a taxonomy and implementations are expected to be able then to drive an experiment just from configuration files [ABJ+22].

The framework could be assessed against different scenarios. First, test-sets may be designed to determine likelihood of failed classification (attack / normal traffic). One of the difficulty is that, to our knowledge, no clear strategy exists to assess the quality of a test set independently of the training procedure. Guidelines exist about how to generate some test set together with a train set to capture a model's performance with respect to scenarios from the state of the art, as done in FREIDA [ABJ+22]. The project could then propose an approach to be as independent as possible from the training procedure during the evaluation of a test set.

References

[ABJ+22] Solayman Ayoubi, Gregory Blanc, Houda Jmila, Thomas Silverston, and Sébastien Tixeuil. Data-driven evaluation of intrusion detectors: a methodological framework. In *International Symposium on Foundations and Practice of Security*, pages 142–157. Springer, 2022.

[AQP+22] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressnegger, Lorenzo Cavallaro, and Konrad Rieck. Dos and don'ts of machine learning in computer security. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 3971–3988, 2022.

[DVV+22] Laurens D'hooge, Miel Verkerken, Bruno Volckaert, Tim Wauters, and Filip De Turck. Establishing the contaminating effect of metadata feature inclusion in machine-learned network intrusion detection models. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 23–41. Springer, 2022.

[HWZ+21] Dongqi Han, Zhiliang Wang, Ying Zhong, Wenqi Chen, Jiahai Yang, Shuqiang Lu, Xingang Shi, and Xia Yin. Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors. *IEEE Journal on Selected Areas in Communications*, 39(8):2632–2647, 2021.

[LEL+22] Lisa Liu, Gints Engelen, Timothy Lynar, Daryl Essam, and Wouter Joosen. Error prevalence in nids datasets: A case study on cic-ids-2017 and cse-cic-ids-2018. In *2022 IEEE Conference on Communications and Network Security (CNS)*, pages 254–262. IEEE, 2022.

[LGH+22] Maxime Lanvin, Pierre-François Gimenez, Yufei Han, Frédéric Majorczyk, Ludovic Mé, and Eric Totel. Errors in the cicids2017 dataset and the significant differences in detection performances it makes. In *International Conference on Risks and Security of Internet and Systems*, pages 18–33. Springer, 2022.

[LTRS21] Thomas Liao, Rohan Taori, Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[PPJ+19] Feargus Pendlebury, Fabio Pierazzi, Roberto Jordaney, Johannes Kinder, and Lorenzo Cavallaro. {TESSERACT}: Eliminating experimental bias in malware classification across space and time. In *28th USENIX security symposium (USENIX Security 19)*, pages 729–746, 2019.

[SLGP23] Mohanad Sarhan, Siamak Layeghy, Marcus Gallagher, and Marius Portmann. From zero-shot machine learning to zero-day attack detection. *International Journal of Information Security*, 22(4):947–959, 2023.