

Study of the Network Intrusion Detection Datasets’ Labels

Gregory Blanc, Benoit Nougnanke, Thomas Robert

September 2026

1 Context

Labeled datasets are central to the training and evaluation of machine learning-based Network Intrusion Detection Systems (NIDS), yet their labeling accuracy is often assumed rather than critically assessed. Indeed, in practice, ground truth is only partially observable, and direct verification of label correctness is rarely possible from the released dataset alone. Existing approaches typically emphasize either product-based noise signals or process-based auditing, but they do not provide a unified and explicit way to reason about label acceptability for a specified use.

In particular, in the face of ongoing research on network traffic generation for the purpose of assessing network intrusion detection systems [PAMP24, VDV⁺26], there is a need to better characterize these generation processes, as many errors, and notably with respect to labels, have been demonstrated in the past [LGH⁺22, LEL⁺22, FEAD24]. These labels have indeed an impact on the performance of the downstream network traffic classification, which have to be taken into account when designing an intrusion detection solution.

2 Proposed work plan

- Review of the literature on labelling processes and issues.
- Study a selection of labelling processed found in the network traffic classification literature (e.g., ConCap [VDV⁺26], Hera [PAMP24], Manatee ¹ or EvidenceForge ²).
- Design a benchmark to assess these different labelling processes incl. testing methodologies and metrics., and their fitness for the network intrusion detection domain.
- Produce a comparative study of a selection of process-based tools.

References

- [FEAD24] Robert Flood, Gints Engelen, David Aspinall, and Lieven Desmet. Bad design smells in benchmark nids datasets. In *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*, pages 658–675. IEEE, 2024.
- [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 Éric 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.

¹Malicious Activity Navigation and Annotation Tool for Event Explanation. Link: <https://gitlab.inria.fr/pirat-public/manatee>

²Realistic Synthetic Logs Generation. Link: <https://github.com/Cisco-Talos/EvidenceForge/>

- [PAMP24] Daniela Pinto, Ivone Amorim, Eva Maia, and Isabel Praça. A novel approach to network traffic analysis: the hera tool. In *2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1850–1856. IEEE, 2024.
- [VDV⁺26] Miel Verkerken, Laurens D’hooge, Bruno Volckaert, Filip De Turck, and Giovanni Apruzzese. Concap: Practical network traffic generation for (ml-and) flow-based intrusion detection systems. In *2026 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 806–828. IEEE, 2026.