# Generalization in ML-based NIDS: Data Diversity, Task Complexity, and Training Dynamics

Gregory Blanc, Benoit Nougnanke, Thomas Robert

***Topics of interest*** – Machine learning, network intrusion detectors, data diversity, classification complexity, training dynamics, generalization

## 1 Context

A Network Intrusion Detection System (NIDS) acts as a *watchdog* over network traffic: it monitors flows, looks for suspicious activity, and raises alerts. Traditionally, this has been done with *signature/rule-based* methods (e.g., matching known attack patterns). Signatures are effective for known threats but struggle with fast-evolving or previously unseen attacks and can be fragile in complex environments. An alternative is *ML-based NIDS*, which casts detection as a learning problem: supervised *classification* (benign vs. attack) and/or *anomaly detection* for out-of-distribution behaviors.

The central challenge is *generalization*: ML-based NIDS models can score well on a benchmark dataset yet fail after deployment when classes are narrow or biased. Such failures often stem from deployment shifts (e.g., new subnets, altered traffic mix, protocol/version updates) not reflected in the benchmark. The key idea is that richer *intra-class diversity*—the variety of behaviors within a benign or attack class—supports more robust decision boundaries, whereas low diversity and lab artifacts inflate scores but do not transfer to real traffic [1, 2]. In such cases, tweaking architectures or hyperparameters rarely fixes the problem: what matters is what the data actually covers and how the learner behaves during training. We therefore need to *characterize* both the data and the learning process. In the broader ML literature (vision, NLP), this is done with diversity metrics (entropy, Vendi Score) [3], classification–complexity measures [4], and training-dynamics data maps that organize examples into easy/ambiguous/hard regions [5]. Adapting these characterization tools to the NIDS context is still an open research perspective.

Building on these observations, the project adopts a data-centric approach for ML-NIDS: we will develop NIDS-adapted measures of intra-class diversity, relate them to task complexity and training dynamics, and quantify how these characteristics are connected to generalization using neural models on standard NIDS benchmarks. The expected outcome is a small, reproducible workbench (training pipeline, assessment, numerical analysis) and a compact dashboard of metrics/plots that make "when and why" a model generalizes more transparent.

**Goal.** Develop NIDS-adapted intra-class diversity metrics and relate them to classification complexity (including training dynamics) to explain and improve ML-NIDS generalization in practice.

## 2 Activities

### 2.1 State of the Art

- Study, compare, and adapt data-centric characterization methods in ML (intra-class diversity, classification complexity, training-dynamics maps) to NIDS data.
- Analyze the ML-based NIDS landscape (datasets, features, architectures, evaluation practices) and pinpoint gaps related to intra-class diversity and generalization.

## 2.2 Research Proposition

- Define and validate intra-class diversity metrics tailored to NIDS, with clear properties (interpretability, comparability, reliability).
- Model links between diversity, task complexity, and training dynamics; formulate testable hypotheses on their predictive power for performance and generalization.

## 2.3 Experimentation

- Build a reproducible pipeline (data preparation, neural training, metric computation, analysis dashboard) implementing the proposed characterization.
- Evaluate on NIDS benchmark datasets [6] and *state-of-the-art* ML-based NIDS; quantify links between these characteristics and performance/generalization (F1, TPR/FPR, AUC), with robustness checks across splits, feature groups, and similarity choices.

# References


[1] B. Nougnanke, G. Blanc, and T. Robert, "How Dataset Diversity Affects Generalization in ML-based NIDS," in *ESORICS 2025 - 30th European Symposium on Research in Computer Security*, Toulouse, France, Sep. 2025. [Online]. Available: https://hal.science/hal-05194730

[2] R. Flood, G. Engelen, D. Aspinall, and L. Desmet, "Bad design smells in benchmark nids datasets," in *2024 IEEE 9th European Symposium on Security and Privacy (EuroS&P)*, July 2024, pp. 658–675.

[3] D. Friedman and A. B. Dieng, "The vendi score: A diversity evaluation metric for machine learning," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=g97OHbQyk1

[4] A. C. Lorena, L. P. Garcia, J. Lehmann, M. C. Souto, and T. K. Ho, "How complex is your classification problem? a survey on measuring classification complexity," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–34, 2019.

[5] S. Swayamdipta, R. Schwartz, N. Lourie, Y. Wang, H. Hajishirzi, N. A. Smith, and Y. Choi, "Dataset cartography: Mapping and diagnosing datasets with training dynamics," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Nov. 2020, pp. 9275–9293.

[6] P. Goldschmidt and D. Chudá, "Network intrusion datasets: A survey, limitations, and recommendations," *Computers & Security*, vol. 156, p. 104510, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404825001993