Evaluation Framework for ML-based IDS

Solayman Ayoubi[1], Gregory Blanc[2], Houda Jmila[2], Thomas Silverston[1], and Sébastien Tixeuil[3]

[1]Université de Lorraine, CNRS, LORIA, Nancy, France
{solayman.ayoubi, thomas.silverston}@loria.fr
[2]SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, Palaiseau, France
{gregory.blanc, houda.jmila}@telecom-sudparis.eu
[3]Sorbonne Université, CNRS, LIP6, Institut Universitaire de France, Paris, France
sebastien.tixeuil@lip6.fr

***Abstract*—Intrusion detection is an important topic in cybersecurity research, but the evaluation methodology has remained stagnant despite advancements including the use of machine learning. In this paper, we design a comprehensive evaluation framework for Machine Learning (ML)-based IDS and take into account the unique aspects of ML algorithms, their strengths, and weaknesses. The framework design is inspired by both i) traditional IDS evaluation methods and ii) recommendations for evaluating ML algorithms in diverse application areas. Data quality being the key to machine learning, we focus on data-driven evaluation by exploring data-related issues.**

***Index Terms*—Intrusion Detection System, Machine Learning, Data-driven Evaluation, Evaluation Framework**

I. Introduction

For over 30 years, most intrusion detection systems have been evaluated in a similar manner, ignoring data-related best practices, even after the introduction of machine-learning approaches. Worse still, fastly-aging datasets were perused for more than 20 years.

In the literature related to ML-based IDS, there is a general focus on measuring the *Attack detection accuracy*, while ignoring other properties. Attack detection is a *measurement methodology* as described by Milenkoski et al. [7], where the classification accuracy of an IDS is measured in the presence of mixed workloads (benign and malicious traffic). Besides, additional ML-related issues, such as data bias that affect the generalization or stability of ML-based IDS, are seldom considered. Therefore, in order to enhance the overall quality of the evaluation, we propose a generic approach to evaluate machine learning-based IDS from multiple perspectives: we go beyond the classical quantitative evaluation methods, that solely focus on measuring effectiveness using fundamental metrics, and consider data-driven evaluations by focusing on the data used for the assessment. We then wish to evaluate concepts more specific to Data-related models like data quality or representativeness.

II. Proposal of an Evaluation Framework

One of the objectives of this framework is to bring together the different evaluation methods found in the literature, in particular those that propose to evaluate aspects specific to the use of machine learning such as robustness and generalization, and to suggest a method for researchers to properly assess their detection models. Our research is inspired by Milenkoski et al. [7], who define the measurement methodology of an evaluation property as the selection of appropriate workload (dataset) and metrics. Our proposal adapts this approach to ML-based IDS and embeds it into a framework that generalizes the evaluation of several properties beyond detection performance (also known as *effectiveness*). Our complete framework can be found in Figure 1. The framework is divided into several modules that contribute to the complete evaluation process. We further detail each module in what follows.

A. Properties

This module allows an evaluator to select a set of properties that the target IDS (system under test) is assessed against.

Effectiveness is the usual property for assessing the detection performance of an IDS. However, relying solely on performance evaluation is one of the major issues in the evaluation of ML-based IDS, since other crucial characteristics, such as the ML algorithm's robustness or generalization must be considered. Besides *effectiveness*, the properties we propose in our framework are influenced both by works in the domain of intrusion detection and data-related problems in ML: i) *effectiveness*, measures how well the IDS detects intrusion; ii) *efficiency* measures how many computing resources the IDS requires; iii) *usability* measures how easy it is for a non-security expert to use the IDS; iv) *actionability* measures how useful are the alerts for a security operator; v) *robustness* measures how well the IDS sustains incidents or attacks directed against it (e.g., adversarial examples, concept drift); vi) *intrusiveness* measures the privacy risks on the data manipulated by the IDS; vii) *collaborativeness* measures how well the system collaborates with other security mechanisms.

B. Datasets

As the main focus of our approach, the dataset module is central in our framework, deriving which datasets are appropriate to evaluate a property, and feeding them to the evaluation module. Indeed, the kind of dataset to be utilized is determined by the requirement to evaluate a specific property.

This work is funded by the GRIFIN project (ANR-20-CE39-0011)

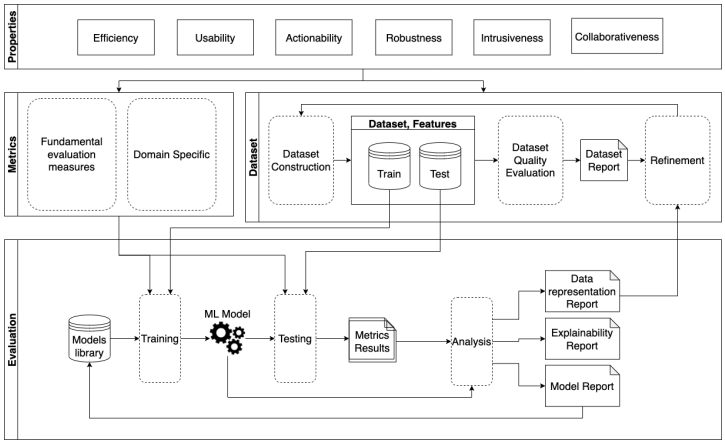

Fig. 1: Data-driven Evaluation framework for ML-based IDS

This module has 3 main processes: *construction*, *evaluation*, and *refinement*.

a) Dataset construction: This process produces one or several datasets (each of them later split into a training set and a test set) that may be represented according to various subsets of features. Similar to Milenkoski et al. [7], we consider various sources of data, ranging from raw traffic captures to extracted flows to packet traces to feature vectors that have been generated from a broad set of environments including production environments (rare!), emulation/simulation testbeds, or legitimate and attack traffic generation tools. Generation tools also encompass generative methods that output synthetic feature vectors. These sources also come as readily exploitable datasets, some of them have been shared among the IDS research community. A dataset may actually enable the evaluation of more than one property.

b) Dataset evaluation: We suggest evaluating the dataset early so that it might potentially be improved through a refinement stage in order to get the best evaluation possible. For example, Gharib et al. [4] have proposed a weighted score using 11 criteria to evaluate the quality of an intrusion detection dataset. Also, Wasielewska et al. [9] propose to experimentally investigate the limits of detection by using their dataset quality assessment method (PerQoDA).

c) Dataset refinement: The goal of the dataset refinement process is to use the various reports from the model evaluation, as well as the dataset evaluation, to raise the dataset's quality. Initially, we can easily address the many issues brought up by the assessment using Gharib's method: for instance, if we discover a deficit in the proportion of attacks, we can try to add the missing traffic. However, after receiving feedback from a first training session, particularly from the *data representation report*. For example, if changes need to be made we can change the dataset's feature set by reducing the dataset's dimensionality.

C. Metrics

In this section, we detail the families of metrics that are needed to produce an accurate and customized evaluation. It is essential to select the appropriate metrics in order to properly analyze a property.

Bekkar et al. [1] expressly identifies three groups of metrics as follows, and we add a fourth specific to the assessment of IDS.

a) Fundamental evaluation measures: This class relates to the metrics that can be calculated using the confusion matrix, including *accuracy*, *precision*, and *recall*.

b) Combined evaluation measures: These metrics combine the fundamental measures in a way that they are less susceptible to potential class imbalance.

c) Graphical performance evaluation: In this category, the metrics are based on the ROC curve: the true positive rate (TPR) and false positive rate (FPR) are plotted against one another at different threshold values.

d) Domain specific: This category of metrics outlines the metrics created expressly for the assessment of IDS, such as the C_{ID} a metric defined by Gu et al. [5] in 2006.

D. Evaluation

This module evaluates a system under test (an IDS) for a given set of properties, and their appropriately derived datasets and metrics. Aside from model training and testing, the subsequent results are analyzed to refine both the model fueling the ML-based IDS and the dataset.

a) Training and Testing: These processes in the evaluation module are the most simple and common ones, yet mandatory. The result of the training process is the trained model and validated model. This model is then used in the testing process (also known as *inference*) to output the *metrics results*, which include the outcomes of the selected metrics computed using the test set. These reports are often found in other publications evaluating IDS proposals using the classical methodology and contain different values of the fundamental metrics for a set of model architectures.

b) Analysis: The incorporation of an analysis process is the foremost improvement we advocate for model evaluation. Through this process, we can acquire several reports that are

highly helpful for both improving the IDS and performing a more comprehensive evaluation.

The *data representation report* evaluates the suitability of our dataset for the model. The initial assessment provides a general quality measure, while the evaluation after testing uses performance metrics. The report determines appropriate features and informs refinement in future iterations. Since some ML (rather Deep Learning) models are often considered black boxes, hindering interpretation of results. To address this, XAI techniques have emerged to provide explanations of model outcomes.

The *explainability report* details the application of such methods to the evaluation of IDS models. The *model report* clarifies whether the chosen model is suitable for the desired task. We may want to assess several models from which we derive the various performance measures. From these outcomes, we produce this report to demonstrate the effectiveness of the employed algorithms. This report allows us to modify the model library's list of models so that we only keep the most effective ones in an evolutionary approach.

In conclusion, the framework provides a methodological blueprint for developing the appropriate dataset and the assessment procedure. Some of the activities are loops that enable the improvement of various evaluation components, such as the dataset and model selection, at each iteration.

III. Related Work

Throughout the years, researchers have presented numerous IDS evaluation approaches. Here, we introduce a few of them and compare them with the proposed framework.

In 2006, Berúmdez-Edo et al. [2] proposed a method based on dataset partitioning. Their method is one of the options for constructing the dataset since it provides a means to prepare the databases for model training, testing, and evaluation. A methodology based on a novel metric that plots all factors affecting an IDS's performance was presented by Cardenas et al. [3] in the same year. This metric is one of the few existing domain-specific metrics. Milenkoski et al. [7] presents a method in 2015 to evaluate several IDS properties, and they characterize the evaluation of a property as a *measurement methodology*, i.e.,the selection of a reliable dataset and adapted metrics. As a result, we can assess more properties if we can specify the appropriate data and metrics. In 2020, Magán-Carrión et al. [6] proposed a methodology that specifies the best practices for pre-processing the dataset, training, and assessing the model. In the end, their approach focuses on standardizing model preparation rather than introducing any new evaluation techniques. The distinction between our proposal and the current methodologies is clearly shown in Table I, where many of the features of our framework are either partially implemented or missing. Indeed, the various evaluation techniques do only consider one aspect at a time.

	Our proposal	**[7]**	**[6]**	**[2]**	**[3]**
Properties	✓	✓	Partially	Partially	Partially
Dataset Construction	✓	✓	Partially	Partially	
Dataset Evaluation	✓				
Refinement	✓				
Domain Specific Metrics	✓	✓			✓
Analysis	✓				

TABLE I: Comparison of our framework with other evaluation methods

IV. Conclusion

We can conclude that, despite improvements in IDS design, evaluation techniques have barely changed. In response to this issue, we propose an evaluation framework which goals are to complete knowledge gaps and standardize evaluation practices. It is obvious from comparing our approach to the state of the art that our proposal includes a greater number of crucial factors that need to be assessed. Our future works focus on developing our approach, including 1) implementing the framework, in particular by specifying the link between a property and its measurement methodology, i.e., the associated datasets and metrics, 2) formalizing the evaluation part of the framework, and 3) constructing a benchmark to evaluate and compare various ML-based intrusion detection systems. A more detailed version of this article appeared in FPS 2022 [8].

References

[1] Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, 3(10), 2013.

[2] María Bermúdez-Edo, Rolando Salazar-Hernández, J Díaz-Verdejo, and Pedro Garcia-Teodoro. Proposals on assessment environments for anomaly-based network intrusion detection systems. In *International Workshop on Critical Information Infrastructures Security*, pages 210–221, 2006.

[3] A.A. Cárdenas, J.S. Baras, and K. Seamon. A framework for the evaluation of intrusion detection systems. In *2006 IEEE Symposium on Security and Privacy (S&P'06)*, pages 15–77, 2006.

[4] Amirhossein Gharib, Iman Sharafaldin, Arash Habibi Lashkari, and Ali A Ghorbani. An evaluation framework for intrusion detection dataset. In *2016 International Conference on Information Science and Security (ICISS)*, pages 1–6. IEEE, 2016.

[5] Guofei Gu, Prahlad Fogla, David Dagon, Wenke Lee, and Boris Skorić. Measuring intrusion detection capability: An information-theoretic approach. In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pages 90–101, 2006.

[6] Roberto Magán-Carrión, Daniel Urda, Ignacio Díaz-Cano, and Bernabé Dorronsoro. Towards a reliable comparison and evaluation of network intrusion detection systems based on machine learning approaches. 10(5):1775. Publisher: Multidisciplinary Digital Publishing Institute.

[7] Aleksandar Milenkoski, Marco Vieira, Samuel Kounev, Alberto Avritzer, and Bryan D Payne. Evaluating computer intrusion detection systems: A survey of common practices. 48(1):1–41. Publisher: ACM New York, NY, USA.

[8] Ayoubi Solayman, Blanc Gregory, Jmila Houda, Thomas Silverston, and Tixeuil Sébastien. Data-driven evaluation of intrusion detectors : a methodological framework. In *15th Symposium on Foundations and Practice of Security*. Springer, 2022.

[9] Katarzyna Wasielewska, Dominik Soukup, Tomáš Čejka, and José Camacho. Evaluation of Detection Limit in Network Dataset Quality Assessment with Permutation Testing. In *4th Workshop on Machine Learning for Cybersecurity (MLCS)*, 2022.