Data-driven evaluation of intrusion detectors: a methodological framework

Solayman Ayoubi, Gregory Blanc, Houda Jmila, Thomas Silverston, Sébastien Tixeuil

Data-driven Evaluation of Intrusion Detectors: a Methodological Framework

Solayman Ayoubi[1][0000−0001−5711−4402], Gregory Blanc[2][0000−0001−8150−6617], Houda Jmila[2][0000−0002−4864−5380], Thomas Silverston[1][0000−0003−0451−5637], and Sébastien Tixeuil[3][0000−0002−0948−7172]

[1] LORIA, Universite de Lorraine, France
[2] SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, France
[3] Sorbonne Université, CNRS, LIP6, Institut Universitaire de France, France

Abstract. Intrusion detection systems are an important domain in cybersecurity research. Countless solutions have been proposed, continuously improving upon one another. Yet, and despite the introduction of distinct approaches, including machine-learning methods, the evaluation methodology has barely evolved.
In this paper, we design a comprehensive evaluation framework for Machine Learning (ML)-based intrusion detection systems (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. Our approach goes beyond evaluating intrusion detection performance (also known as *effectiveness*) and aims at proposing standard data manipulation methods to tackle robustness and stability. Finally, we evaluate our framework through a qualitative comparison with other IDS evaluation approaches from the state of the art.

Keywords: Intrusion Detection System · Machine learning · Data-driven Evaluation · Evaluation Framework

1 Introduction

It has been almost twenty years since the publication of the NIST internal report on testing intrusion detection systems [29]. The NIST report identified 10 measurable characteristics, and 4 challenges (incl. how to use background traffic to test IDS), and presented recommendations to improve both datasets and metrics. While some of these characteristics and challenges remain relevant, they also highlight the need to update and improve our IDS evaluation approaches.

Although new techniques like artificial intelligence were introduced to intrusion detection systems, researchers still use outdated evaluation methodologies and datasets. Since 2006, the article by Bermúdez-Edo et al. [8] revealed that the databases used for IDS evaluation are obsolete, however, they are still used

today. According to Tavallaee et al. [37], in 2010, almost 28% (resp. 24%) of research papers used the obsolete KDD99 (resp. DARPA) datasets.

Milenkoski et al. [30] proposed an evaluation technique based on a design space comprised of a workload (dataset property), customized metrics, and a measurement methodology. In their publications, Milenkoski et al. suggest a number of measurement methodologies that correlate to the potential property (Attack detection accuracy, Resistance to evasion techniques...) for evaluation.

In the literature related to ML-based IDS, which generally focuses on the property *Attack detection accuracy*, which is essentially a measurement methodology as described by Milenkoski et al.and is defined as the accuracy of an IDS in the presence of mixed workloads (benign and malicious traffic), *Resistance to evasion methods* or *Resource consumption* are rarely covered. However, additional ML-related issues, such as the bias from the data, also affect the generalization or stability of the ML-based IDS.

Furthermore, despite the datasets' obvious quality problems, they are nonetheless used without any oversight. Therefore, in order to enhance the overall quality of the evaluation, we propose a generic and general 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 considers data-driven evaluations by focusing on the data used for the assessment. In the IDS context, we analyze machine-learning concerns like explainability and robustness to adversarial examples. To do so, we examine (i) IDS-specific assessment methods, (ii) AI-specific evaluation methods that can be applied to IDS, and (iii) relevant recommendations from the state of the art [5] and the standards [29].

This article is structured as follows: Sec. 2 presents some related works, Sec. 3 analyses a number of IDS solutions with a focus on evaluation methods. Then, we present our proposal in Sec. 4 as well as a generic evaluation framework. Finally, we conclude in Sec. 5.

2 Related Work

Throughout the years, researchers have presented numerous IDS evaluation approaches. In this part, we introduce some of them. All of the methodologies make an effort to give researchers resources to assess IDS.

Milenkoski et al. [30] identify the most common practices to evaluate different types of intrusion detection systems. To do so, they define a three-part design space including (i) workloads, which are testing sets of data, and their means of production; (ii) metrics, which quantify performance-related properties (non-functional with respect to IDS), or security-related ones; (iii) measurement methodology, which specifies the evaluation properties along with its associated workloads and metrics.

Indeed, they include methods and tools to generate workloads and focus on metrics that quantify the accuracy of the detection. Our proposed framework is inspired by the measurement methodology proposed by Milenkoski et al. but

shifts its initial paradigm towards machine-learning-based IDS, i.e., it relies on the evaluation of best practices from the field of machine learning, in particular with respect to data-related issues.

Magán-Carrión et al. [26] examine Network IDS (NIDS) solutions and point out the lack of a standardized method for evaluating machine learning-based NIDS. According to the authors, it is challenging to compare various NIDS because the state of the art does not provide enough information on the evaluation methods. Hence, their methodology 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, they clearly present the different training stages of a model: Feature Engineering, Feature Selection, Data Pre-processing, Hyper-parameters Selection, and Performance Metrics.

Bermúdez-Edo et al. [8] suggest requirements for implementing standardized IDS evaluation framework. The authors present a new method for evaluating anomaly-based IDS with a focus on data-partitioning approaches. The authors then offer a technique to get the databases ready for model training, testing, and evaluation. They outline 3 steps: 1. they separate the attacks in one set and the normal in another set, 2. they split the two datasets between a training set, a test set, and a validation set, and 3. they combine some parts to produce three final datasets (train, test, validation). In this method, the authors concentrate on dataset partitioning.

Cardenas et al. [10] presents an IDS evaluation framework that allows for a consistent comparison of the most used metrics in the literature. The authors present a graphical method for comparing the different metrics for a wide range of parameters. They provide a new metric that plots all variables influencing an IDS performance. According to the authors, it is more interesting to determine the IDS that performs best against the most severe attacks than on average. The proposed metric is beneficial for our approach since it enables the results of other domain-specific metrics to be summarized.

3 Analysis of Evaluation Approaches in ML-based IDS

In this section, we review evaluation methods employed in recent ML-based IDS publications in order to identify common practices that help create a generic evaluation approach. We selected the publications from recent surveys [2, 11] which respectively presented articles from 2019-2021 and 2015-2018 and updated the list with new papers. Following many searches using terms such as "intrusion detection" or "ML-based intrusion detection", we retained the most recent publications (2020–2022) that fell under the scope.

Table 1 highlights a few components of the evaluation method employed in these publications, namely the dataset and the metrics used, as well as some specific evaluation measures beyond what could be described as a common evaluation approach. The remainder of this section details the various evaluation

measures that we have noticed in this corpus of publications, both common (*classical measures*) and specific to each publication.

Classical measures. From our survey of the state of the art in machine-learning-based (network) IDS, the evaluation measures employed by the researchers rarely differ. Although we did find more peculiar measures (as detailed in the last columns of Table 1), a common, allegedly conventional, methodology stood out. This classical evaluation can be defined using the methods introduced by Magán-Carrión et al. [26].

Some examined publications [13, 22, 35, 38] that fall into this category solely advocate for obtaining and contrasting the outcomes of basic metrics (accuracy, precision, and recall) on various model architectures. For instance, in order to enhance the performance of their deep neural networks, the writers of these publications compared several architectures by varying some parameters such as the size of the hidden layers for Gao et al. [13], the neural network activation functions for Thing [38], the number of memory blocks and cells in LSTM for Staudemeyer [35], and finally, the learning rate and the size of the hidden layers for Kim et al. al. [22].

Data-related measures. Data-related measures encompass any evaluation techniques dealing with data-related manipulation, e.g., augmenting the dataset, reducing its dimensionality, generating data with a specific environment, and random resampling. We are primarily interested in these methods given that we wish to evaluate ML-based IDS.

Zhang et al. [44] leverage SMOTE to create the missing data in the unbalanced NSL-KDD dataset. This results in increasing the detection performance of their CNN-based IDS on previously under-represented classes. Tang et al. [36] heavily reduced the data representation of the NSL-KDD dataset from 41 features to 6. This makes their DNN-based flow anomaly detector more efficient. Zolotukhin et al. [46] used the Realistic Global Cyber Environment (RGCE) to run their simulation, RGCE is a closed environment that replicates the user traffic and organizational structures of the real Internet. This article is included in our survey's environment category since it makes use of a simulated environment. Al-Qatf et al. [3] suggest combining SVM and Sparse Autoencoder. The following two methods are used to assess the effectiveness of their method using the NSL-KDD dataset, for this purpose a ten-fold cross-validation is carried out for both training and testing. Random resampling can be done using the k-fold cross-validation method.

Multi-label measures. ML-based IDS are often termed behavioral IDS or anomaly-based IDS[4], that is, binary classifiers attempting to distinguish malicious traffic from a normal one. But some datasets offer more depth in exhibiting several

[4] We believe however that the term "anomaly-based IDS" should solely apply to IDS trained on normal traffic only.

	UNSW	Private	CICIDS	KDD99	Kyoto-Honeypot	Others	NSL-KDD		Accuracy	Precision / Recall	F-measure	ROC Curve / AUC	Confusion matrix / FAR	Anomaly Score (AS)	ERR	DR	EIR	MCC	Squared reconstruction error	Specificity	Efficiency		Classification tasks ♦	Environment □	Dataset manipulation □	Resampling test set □	Generate data □	Multi-label evaluation ♦	Model architecture §	
[43]						✓			✓	✓	✓	✓	✓										✓							
[38]						✓			✓																				✓	
[46]						✓			✓			✓	✓											✓						
[14]						✓								✓											✓					
[4]				✓					✓				✓																✓	
[36]							✓		✓	✓	✓														✓					
[3]							✓		✓	✓	✓												✓			✓				
[31]		✓										✓	✓		✓									✓						
[6]							✓		✓	✓	✓														✓	✓				
[25]							✓		✓	✓	✓		✓															✓		
[24]							✓									✓	✓										✓			
[34]				✓						✓	✓																✓			
[45]	✓								✓	✓	✓																✓			
[42]						✓			✓	✓	✓		✓														✓			
[20]						✓	✓					✓			✓												✓			
[27]					✓		✓			✓	✓							✓							✓	✓				
[23]					✓	✓	✓				✓							✓							✓	✓			✓	
[13]				✓					✓						✓				✓										✓	
[35]						✓			✓	✓															✓				✓	
[1]			✓						✓														✓		✓			✓		
[28]		✓							✓	✓	✓	✓												✓						
[21]			✓							✓	✓		✓		✓											✓				
[19]							✓			✓	✓		✓							✓					✓	✓				
[44]							✓		✓						✓												✓	✓		
[22]				✓					✓						✓						✓								✓	
	Dataset								Metrics														Approaches							

Table 1: Comparison of the surveyed publications. Most IDS evaluations employ a subset of the above datasets and metrics with little to no variation. Additional measures deal with varying model architectures (§), multi-label classification (♦) or data-related manipulations (□)

classes of attacks, which could be interesting for IDS to discriminate with respect to producing a specific intrusion response. To that end, multi-label classification is employed. We have found some works measuring its advantage, either in comparison with binary classification or in evaluating per-class performance.

For example, Yu et al. [43] propose a novel network intrusion model by stacking dilated convolutional autoencoders and they evaluate their method on two new intrusion detection datasets. Several experiments were carried out to check the effectiveness of their approach. They used two different datasets: CTU-UNB & Contagio-CTU-UNB and six classical evaluation metrics. To evaluate they perform 3 types of classification tasks: 6-class classification using the Contagio-

CTU-UNB dataset and 2-class and 8-class classification using the CTU-UNB dataset.

Moreover, Abbas et al. [1] proposed an ensemble model combining Naive Bayes, Logistic Regression, and a Decision Tree. In order to assess the performance of their suggestion they determine the accuracy of their model for each label. They end up with a total of 15 different accuracies, each of which represents the detection performance for this label.

Table 1 compares ML-based IDS proposals with respect to their evaluation methodology. What can be observed is that they often share the same evaluation approach. Many evaluations were replicating approaches previously seen in the state of the art, and the trend has been shifting over the years, for example from computing accuracy only to computing both precision and recall instead. It is still the case today although intrusion detection-specific metrics were proposed [10, 17, 18, 39]. Another worrying aspect is the choice of the dataset. Although NSL-KDD has been perused for many years, many datasets were created and shared in the last 10 years. It affects evaluation in its timeliness as the attacks it contains are outdated and far from the sophistication of modern attacks. Often, other simple data-related issues, e.g., unbalance, are addressed using supplementary evaluation measures such as augmenting the dataset or reducing its dimensionality.

Finally, additional measures that we have observed with respect to multi-label classification, dataset construction, or model architectures are seldom used in combination, reducing the quality of the models trained and tested. This advocates for the definition and formalization of a holistic framework enabling researchers of the domain in mastering the ML pipeline and adapting it to the task of evaluating ML-based IDS with respect to a wide range of properties including detection performance and resource consumption, of course, but also generalization, robustness, and so on.

4 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 models. Our research is inspired by Milenkoski et al. [30], 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*). In particular, it focuses on a dataset construction component as a generalization of the workload concept and extends it to accommodate feedback from the evaluation analysis, ultimately providing continuous improvement to both the ML models used by the IDS and the data representation they use. Not only does the property have an impact on the

metrics that evaluate it, but the dataset may embed some challenges that the metrics should account for (*e.g.*, when using unbalanced datasets).

We also want to add some aspects not yet studied enough in the application of machine learning to IDS such as explainability. The complete framework can be found in Figure 1. The framework is divided into several modules that contribute to the complete evaluation process. The first module focuses on the property that we want to examine. From the selected properties, the *metrics module* will output a set of relevant metrics, and the *dataset module* will construct the appropriate dataset to assess them. Both outputs form the *experiment setting* that will configure the *evaluation module* which will perform the training and testing of one or several models to be assessed by the evaluator. We further detail each module in the ensuing subsections.

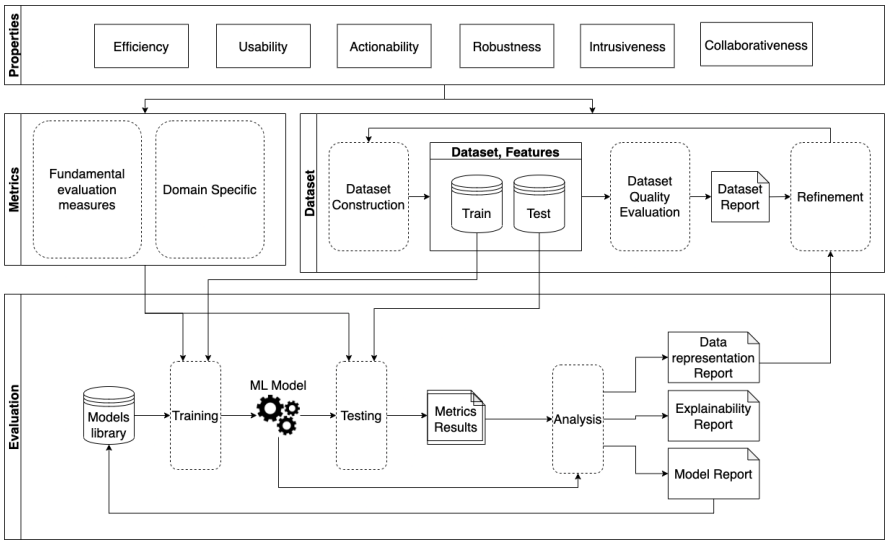

Fig. 1: Data-driven Evaluation framework for ML-based IDS

4.1 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 generalizability must be considered.

Besides *effectiveness*, the properties we propose in our framework are influenced by both works in the domain of intrusion detection, such as Axelsson's [5], and data-related problems in ML:

i) *efficiency* measures how many computing resources the IDS requires; ii) *usability* measures how easy it is for a non-security expert to use the IDS; iii) *actionability* measures how useful are the alerts for a security operator; iv) *robustness* measures how well the IDS resists incidents or attacks directed against it (e.g., adversarial examples, concept drift); v) *intrusiveness* measures the privacy risks on the data manipulated by the IDS; vi) *collaborativeness* measures how well the system collaborates with other security mechanisms.

4.2 Datasets

As the main focus of our approach, the dataset module is central in our framework, deriving the datasets 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 module has 3 main processes: *construction*, *evaluation*, and *refinement*.

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. [30], we consider various sources of the 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 comprehensive list of the publicly available datasets that are commonly used is presented by Ring et al.in their survey [32].

Dataset construction outputs datasets that fit the measurement methodology as expressed by Milenkoski et al. [30], that is it enables the evaluation of a given property. A dataset may actually enable the evaluation of more than one property.

For example, Bermúdez-Edo et al. [8] propose steps to acquire and partition a network traffic dataset for evaluating the effectiveness of anomaly-based IDS, among others. Some generation criteria are as follows: i) both *normal* and attack traffic should be present, and the dataset should be partitioned between training (only normal), validation, and test (both types of traffic) sets with realistic proportions; ii) the dataset should be sufficiently voluminous as to be representative of most traffic behaviors. They also described approaches to tackle a number of issues: i) generating *anomalous* traffic (i.e., new attacks in a hybrid setting) by using two filters (one obsolete and one up-to-date), ii) improving the *robustness* of the dataset (increase the effective size of the data available for training and testing) when its size is modest by resampling training and test sets, and averaging the performance results, and iii) updating models by shifting the datasets: the up-to-date rules become the out-of-date rules and the new rules become the up-to-date rules. .

Combining datasets can help to fill in any gaps that may exist in the chosen or generated datasets, eventually creating more *representative* datasets. To be representative the dataset needs samples large enough to adequately reflect the general community's norms, including both permitted and prohibited behaviors.

Regardless of the sources of data, a dataset D^p_{Fj} is an instance of a dataset D^p constructed so as to contain diverse samples allowing the evaluation of a property p, and in which each sample is represented by the set of features F_j. Its split between the training set $Tr(D^i_{F_j})$ and the test set $Tt(D^i_{F_j})$ is conditioned by both the property to be evaluated and the type of IDS (binary classifier, multi-classifier, anomaly detector).

Dataset evaluation. We suggest evaluating the dataset upstream so that it might potentially be improved through a refinement stage in order to get the best evaluation possible.

For example, Gharib et al. [15] have proposed a weighted score on 11 criteria to evaluate the quality of an intrusion detection dataset. The 11 criteria are attack diversity, anonymity, available protocols, complete capture, complete interaction, complete network configuration, complete traffic, feature set, heterogeneity, labeled dataset, and metadata. Practitioners are invited to define weights themselves, that best suit their requirements.

Viegas et al. [40] tackled the issue of realistic network conditions for the evaluation of intrusion detectors by generating datasets using a honeypot with a client-server approach. The generated datasets should satisfy a number of expected properties [40]: i) realism: the produced network traffic can be observed in production environments; ii) validity: packets are well-formed and follow the client-server communication paradigm; iii) prior labeling: samples are correctly labeled to enable correct classification; iv) high variability (diversity): the dataset should present a diverse set of services, client behaviors, and attacks; v) correct implementation: attacks follow a well-known or *de facto* standard; vi) ease of updating: the dataset should incorporate new services and attacks; vii) reproducibility: experts should be able to compare datasets; viii) without sensitive data: the dataset should not contain or reveal sensitive information, so as to be shared among researchers.

Additionally, one might desire more focused techniques, such as evaluating datasets produced by a Generative Adversarial Network. Early works in other fields have emerged, such as the one from Gonçalves et al. [16] that proposed a method for the generation and evaluation of synthetic patient data. Using a set of complementary metrics, they evaluated the quality of the synthetic data generators. These metrics can be divided into 2 groups, the *data utility* and the *information disclosure*. The *data utility* metrics measure how well the synthetic dataset incorporates the statistical characteristics of the original data, and the *information disclosure* metrics quantify to which extent the synthetic data may reveal the real data. They proposed 5 data utility metrics (Kullback- Leibler (KL) divergence, pairwise correlation difference, log-cluster, support coverage, and cross-classification) and 2 information disclosure metrics (membership disclosure, attribute disclosure).

Although the work is not in the field of intrusion detection, it shows promises for evaluating synthetic discrete tabular data that appears frequently in network traffic datasets.

Finally, Wasielewska et al. [41] propose to experimentally investigate the limits of detection by using their dataset quality assessment method (PerQoDA). This method makes it simple to determine whether the dataset's information is comprehensive enough to reliably classify observations. In multidimensional datasets, it can spot irregularities in the connections between observations and labels. An efficient method for evaluating dataset quality aids in understanding how performance outcomes are affected by dataset quality and can be useful in resolving issues relating to the deterioration of model performance. Prior to any ML application, they recommend, assessing the dataset quality.

Dataset refinement. The dataset refinement process describes the step where we use all the observations made to improve the dataset. The goal 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*, one could wish to make adjustments. In this scenario, a variety of strategies can be used to change the dataset's feature set. For example, Bronzino et al. [9] propose a complete method named Traffic Refinery. This approach aims to transform the traffic in real-time to produce a variety of feature representations for machine learning models. With this tool, we can explore and evaluate which representations work best for the property to be evaluated.

Indeed, there is no standard set of features for Network Intrusion Detection Datasets. Different representations may actually yield different performances for the same model. To prove this, Sarhan et al. [33] proposes to evaluate and compare two different sets of features, the Netflow-based features, and the CICFlowMeter features. The evaluation has been conducted on three datasets and using two machine learning classifiers. The results show a constant superiority of the NetFlow features. In addition to this, the authors used SHAP to explain the prediction results of the ML models to identify the key features for each dataset. With this approach, we can choose the best data representation methods.

4.3 Metrics

In this section, we detail the families of metrics that are needed to produce an accurate and customized evaluation, it's essential to pick the appropriate metrics in order to properly analyze a property. Although the metrics described in this part are primarily concerned with detection performance, choosing a dataset and metric based on the evaluation of a property allows for the study of more properties with the same metrics than only effectiveness.

Bekkar et al. [7] expressly identify three groups: *fundamental evaluation measures*, *combined evaluation measures*, and *graphical performance evaluation*. The authors apply these metrics to compute the effectiveness of an IDS in the presence of unbalanced datasets. They remark that accuracy places more weight on the most common classes than on the rare ones so using metrics like accuracy completely skews the results. It appears therefore that one should carefully choose metrics that compensate for a dataset's shortcomings. Even though the authors, in this case, are interested in unbalanced datasets, we recommend using at least the categories of metrics established by Bekkar et al.in order to account for the various defaults of the datasets. The metrics categories that we advise are the following.

Fundamental evaluation measures. This class of metrics relates to the metrics that can be calculated using the confusion matrix's results. Identified fundamental measures include *accuracy*, *precision*, and *recall*.

Combined evaluation measures. The metrics derived from fundamental measures are included in this category. The following metrics can be found: G-means, the likelihood ratios, Discriminant power, F-Measure, Balanced Accuracy, Youden index, and finally the Matthews correlation coefficient (MCC). These metrics combine the fundamental measures in a way that they are less susceptible to potential class imbalance.

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.

The AUC, which is defined as a summary indication of the ROC curve performance, is used to indicate the performance of a classifier into a single measure. But there are also several other metrics such as Weighted AUC, Cumulative Gains Curve and lift chart and Area Under Lift. These metrics provide a concise summary of the fundamental evaluation measures and enable the selection of potentially optimal models while disqualifying subpar ones regardless of the cost context or class distribution.

Domain specific. As early as 2006, Gu et al. [17] employed information theory to model the capability of an IDS to correctly classify normal and intrusive traffic. Their objective was to incorporate existing metrics while not relying on subjective measures and reduce the *uncertainty* about the input given the IDS output. The proposed metric called the *Intrusion Detection Capability*, or C_{ID}, is the ratio of the mutual information between the IDS input and output to the entropy of the input. Mutual information measures the amount of uncertainty of the input resolved by knowing the IDS output. Later, Imoize et al. [18] extended C_{ID} to select an optimal operating point, calculate the expected cost and compare intrusion detectors. To that end, they incorporated a decision-tree-based analysis to determine the optimal operating point, as done by Ulvila and Gaffney [39].

These metrics are some examples of what we can find in the literature to specifically evaluate IDS.

4.4 Evaluation

This module performs the evaluation of 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.

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.

Analysis. The incorporation of an analysis process is the real improvement we advocate for model evaluation. Through this process, we are able to acquire a number of reports that are highly helpful for both the IDS's improvement and its comprehensive evaluation.

The *data representation report* helps determine whether or not our dataset is suitable for the model. Although the initial assessment of the dataset during the construction phase gives us a general quality measure, the evaluation following the test phase enables us to evaluate, using performance metrics, its suitability for our purposes. We obviously want to determine whether a set of features is appropriate for our models. The findings in this report can then be applied to the refinement process in a subsequent iteration of the evaluation.

Since some ML (rather Deep Learning) models are regarded as black boxes that do not allow for a straightforward explanation of their decisions, it impairs the user's ability to interpret the findings. A growing number of techniques known as XAI that enable an explanation of the outcomes have been developed in response to this issue, Charmet et al. [12] conduct a thorough literature review on the connection between cybersecurity and XAI. 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. In fact, we may want to assess a number of models for which we derive the various performance measures. From these outcomes, we produce this report with the aim of demonstrating 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 instructions for developing the assessment environment and 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.

4.5 Qualitative Assessment of the Proposed Framework.

Reference	Properties	Dataset Construction	Dataset Evaluation	Refinement	Domain Specific Metrics	Analysis
Our proposal	✓	✓	✓	✓	✓	✓
[30]	✓	✓			✓	
[26]	Partially	Partially				
[8]	Partially	Partially				
[10]	Partially				✓	

Table 2: Comparison of our framework with other evaluation methods

We outlined current issues with IDS assessment in Section 1 in addition to the fact that many relatively recent works still use outdated evaluation environments. The approaches are obsolete and not designed for the assessment of ML models. In Section 3, we looked for some unique assessment techniques in various intrusion detection proposals. In this section, we offer a qualitative assessment of our suggested framework by contrasting our procedures with those used by other researchers in the literature.

Here, we contrast our suggestion with the articles listed in the section 2. The discrepancies between our proposal and the current methodologies are clearly shown in Table 2, where many of the elements in our framework are either partially or missing. Indeed, the various evaluation techniques do only consider one aspect at a time. For instance, Cardenas et al. [10], and Milenkoski [30] recommend using domain-specific metrics, yet do not recommend studying the model explicability, or evaluating the dataset itself, two features we include in our framework.

5 Conclusion

We observed that relatively few evaluation techniques in the literature include all required elements for a thorough evaluation of ML-based intrusion detection systems, including in particular: dataset evaluation, explainability, etc. As a result, we propose a methodological framework to assess ML-based IDS in a systematic manner. Our framework is constructed as follows: The framework's first module outlines the various properties that we wish to assess, and it links to the metrics and datasets modules. In our framework, we take into account that the metrics and dataset are defined depending on the property to be evaluated. Given that both components are crucial to the assessment process, the metrics module and the dataset module are connected to the last module, the evaluation module. Some of our modules include loops that can be used for fine-tuning specific assessment processes in future iterations of the evaluation.

Our framework paves the way for future research developments, including 1. actually implementing the framework, 2. formalizing the evaluation part of the framework, and 3. construct a benchmark to evaluate and compare various ML-based intrusion detection systems.

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

References

1. Abbas, A., Khan, M.A., Latif, S., Ajaz, M., Shah, A.A., Ahmad, J.: A new ensemble-based intrusion detection system for internet of things. Arabian Journal for Science and Engineering **47**(2), 1805–1819 (2022),
2. Abdelmoumin, G., Whitaker, J., Rawat, D.B., Rahman, A.: A survey on data-driven learning for intelligent network intrusion detection systems. Electronics **11**(2) (2022)
3. Al-Qatf, M., Lasheng, Y., Al-Habib, M., Al-Sabahi, K.: Deep learning approach combining sparse autoencoder with svm for network intrusion detection. IEEE Access **6**, 52843–52856 (2018)
4. Alrawashdeh, K., Purdy, C.: Toward an online anomaly intrusion detection system based on deep learning. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 195–200 (2016)
5. Axelsson, S.: The base-rate fallacy and the difficulty of intrusion detection. ACM Transactions on Information and System Security (TISSEC) **3**(3), 186–205 (2000)
6. Aygun, R.C., Yavuz, A.G.: Network anomaly detection with stochastically improved autoencoder based models. In: 2017 IEEE 4th International Conference on Cyber Security and Cloud Computing (CSCloud). pp. 193–198 (2017)
7. Bekkar, M., Djemaa, H.K., Alitouche, T.A.: Evaluation measures for models assessment over imbalanced data sets. J Inf Eng Appl **3**(10) (2013)
8. Bermúdez-Edo, M., Salazar-Hernández, R., Díaz-Verdejo, J., Garcia-Teodoro, P.: Proposals on assessment environments for anomaly-based network intrusion detection systems. In: International Workshop on Critical Information Infrastructures Security. pp. 210–221 (2006)
9. Bronzino, F., Schmitt, P., Ayoubi, S., Kim, H., Teixeira, R.C., Feamster, N.: Traffic refinery. Proceedings of the ACM on Measurement and Analysis of Computing Systems **5**, 1–24 (2021)
10. Cárdenas, A., Baras, J., Seamon, K.: A framework for the evaluation of intrusion detection systems. In: 2006 IEEE Symposium on Security and Privacy (S&P'06). pp. 15–77 (2006)
11. Chalapathy, R., Chawla, S.: Deep learning for anomaly detection: A survey
12. Charmet, F., Tanuwidjaja, H.C., Ayoubi, S., Gimenez, P.F., Han, Y., Jmila, H., Blanc, G., Takahashi, T., Zhang, Z.: Explainable artificial intelligence for cybersecurity: a literature survey. Annals of Telecommunications (Oct 2022). https://doi.org/10.1007/s12243-022-00926-7, https://doi.org/10.1007/s12243-022-00926-7
13. Gao, N., Gao, L., Gao, Q., Wang, H.: An intrusion detection model based on deep belief networks. In: 2014 Second International Conference on Advanced Cloud and Big Data. pp. 247–252 (2014)

14. García Cordero, C., Hauke, S., Mühlhäuser, M., Fischer, M.: Analyzing flow-based anomaly intrusion detection using replicator neural networks. In: 2016 14th Annual Conference on Privacy, Security and Trust (PST). pp. 317–324 (2016)
15. Gharib, A., Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A.: An evaluation framework for intrusion detection dataset. In: 2016 International Conference on Information Science and Security (ICISS). pp. 1–6. IEEE (2016)
16. Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., Sales, A.P.: Generation and evaluation of synthetic patient data. BMC Medical Research Methodology **20**(1), 108 (May 2020)
17. Gu, G., Fogla, P., Dagon, D., Lee, W., Skorić, B.: Measuring intrusion detection capability: An information-theoretic approach. In: Proceedings of the 2006 ACM Symposium on Information, computer and communications security. pp. 90–101 (2006)
18. Imoize, A.L., Oyedare, T., Otuokere, M.E., Shetty, S.: Software intrusion detection evaluation system: a cost-based evaluation of intrusion detection capability. Communications and Network **10**(4) (2018)
19. Imrana, Y., Xiang, Y., Ali, L., Abdul-Rauf, Z., Hu, Y.C., Kadry, S., Lim, S.: chi;2-bidlstm: A feature driven intrusion detection system based on chi;2 statistical model and bidirectional lstm. Sensors **22**(5) (2022)
20. Intrator, Y., Katz, G., Shabtai, A.: Mdgan: Boosting anomaly detection using multi-discriminator generative adversarial networks. ArXiv **abs/1810.05221** (2018)
21. Khan, M.A.: Hcrnnids: Hybrid convolutional recurrent neural network-based network intrusion detection system. Processes **9**(5) (2021)
22. Kim, J., Kim, J., Thi Thu, H.L., Kim, H.: Long short term memory recurrent neural network classifier for intrusion detection. In: 2016 International Conference on Platform Technology and Service (PlatCon). pp. 1–5 (2016)
23. Kwon, D., Natarajan, K., Suh, S.C., Kim, H., Kim, J.: An empirical study on network anomaly detection using convolutional neural networks. In: 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). pp. 1595–1598 (2018)
24. Lin, Z., yu Shi, Y., Xue, Z.: Idsgan: Generative adversarial networks for attack generation against intrusion detection. ArXiv **abs/1809.02077** (2018)
25. Lopez-Martin, M., Carro, B., Sanchez-Esguevillas, A., Lloret, J.: Conditional variational autoencoder for prediction and feature recovery applied to intrusion detection in iot. Sensors **17**(9) (2017)
26. Magán-Carrión, R., Urda, D., Díaz-Cano, I., Dorronsoro, B.: Towards a reliable comparison and evaluation of network intrusion detection systems based on machine learning approaches **10**(5), 1775, publisher: Multidisciplinary Digital Publishing Institute
27. Malaiya, R.K., Kwon, D., Kim, J., Suh, S.C., Kim, H., Kim, I.: An empirical evaluation of deep learning for network anomaly detection. In: 2018 International Conference on Computing, Networking and Communications (ICNC). pp. 893–898 (2018)
28. Mehedi, S.T., Anwar, A., Rahman, Z., Ahmed, K., Rafiqul, I.: Dependable intrusion detection system for iot: A deep transfer learning-based approach. IEEE Transactions on Industrial Informatics pp. 1–1 (2022)
29. Mell, P., Lippmann, R., Chung, Haines, J., Zissman, M.: An overview of issues in testing intrusion detection systems (2003-07-11 2003)

30. Milenkoski, A., Vieira, M., Kounev, S., Avritzer, A., Payne, B.D.: Evaluating computer intrusion detection systems: A survey of common practices **48**(1), 1–41, publisher: ACM New York, NY, USA
31. Mirsky, Y., autoencoders for online network intrusion detection. ArXiv **abs/1802.09089** (2018)
32. Ring, M., Wunderlich, S., Scheuring, D., Landes, D., Hotho, A.: A survey of network-based intrusion detection data sets **86**, 147–167, publisher: Elsevier
33. Sarhan, M., Layeghy, S., Portmann, M.: Evaluating standard feature sets towards increased generalisability and explainability of ml-based network intrusion detection (2021)
34. Shahriar, M.H., Haque, N.I., Rahman, M.A., Alonso, M.: G-ids: Generative adversarial networks assisted intrusion detection system. 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC) pp. 376–385 (2020)
35. Staudemeyer, R.C.: Applying long short-term memory recurrent neural networks to intrusion detection. South African Computer Journal **56**, 136–154 (2015)
36. Tang, T.A., Mhamdi, L., McLernon, D., Zaidi, S.A.R., Ghogho, M.: Deep learning approach for network intrusion detection in software defined networking. In: 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM). pp. 258–263 (2016)
37. Tavallaee, M., Stakhanova, N., Ghorbani, A.A.: Toward credible evaluation of anomaly-based intrusion-detection methods **40**(5), 516–524
38. Thing, V.L.L.: Ieee 802.11 network anomaly detection and attack classification: A deep learning approach. In: 2017 IEEE Wireless Communications and Networking Conference (WCNC). pp. 1–6 (2017)
39. Ulvila, J.W., Gaffney Jr, J.E.: Evaluation of intrusion detection systems. Journal of Research of the National Institute of Standards and Technology **108**(6), 453 (2003)
40. Viegas, E.K., Santin, A.O., Oliveira, L.S.: Toward a reliable anomaly-based intrusion detection in real-world environments. Computer Networks **127**, 200–216 (2017)
41. Wasielewska, K., Soukup, D., Čejka, T., Camacho, J.: Evaluation of Detection Limit in Network Dataset Quality Assessment with Permutation Testing. In: 4th Workshop on Machine Learning for Cybersecurity (MLCS) (2022)
42. Yin, C., Zhu, Y., Liu, S., Fei, J., Zhang, H.: An enhancing framework for botnet detection using generative adversarial networks. In: 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD). pp. 228–234 (2018)
43. Yu, Y., Long, J., Cai, Z.: Network intrusion detection through stacking dilated convolutional autoencoders. Security and Communication Networks **2017**, 4184196 (Nov 2017)
44. Zhang, X., Ran, J., Mi, J.: An intrusion detection system based on convolutional neural network for imbalanced network traffic. In: 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT). pp. 456–460 (2019)
45. Zixu, T., Liyanage, K.S.K., Gurusamy, M.: Generative adversarial network and auto encoder based anomaly detection in distributed iot networks. In: GLOBECOM 2020 - 2020 IEEE Global Communications Conference. pp. 1–7 (2020)
46. Zolotukhin, M., Hämäläinen, T., Kokkonen, T., Siltanen, J.: Increasing web service availability by detecting application-layer ddos attacks in encrypted traffic. In: 2016 23rd International Conference on Telecommunications (ICT). pp. 1–6 (2016)