Fall 9-28-2016

Are Intrusion Detection Studies Evaluated Consistently? A Systematic Literature Review

Nuthan Munaiah
Rochester Institute of Technology

Andrew Meneely
Rochester Institute of Technology

Ryan Wilson
Rochester Institute of Technology

Benjamin Short
Rochester Institute of Technology

Are Intrusion Detection Studies Evaluated Consistently? A Systematic Literature Review

Preliminary Technical Report

Nuthan Munaiah, Andrew Meneely, Ryan Wilson, Benjamin Short
Department of Software Engineering
Rochester Institute of Technology
Rochester, New York, USA
nm6061@rit.edu, axmvse@rit.edu, rww9008@rit.edu, bs9053@rit.edu

September 28, 2016

Are Intrusion Detection Studies Evaluated Consistently? A Systematic Literature Review

Nuthan Munaiah, Andrew Meneely, Ryan Wilson, Benjamin Short
Department of Software Engineering
Rochester Institute of Technology
Rochester, New York, USA
nm6061@rit.edu, axmvse@rit.edu, rww9008@rit.edu, bs9053@rit.edu

September 28, 2016

Abstract

Cyberinfrastructure is increasingly becoming target of a wide spectrum of attacks from Denial of Service to large-scale defacement of the digital presence of an organization. Intrusion Detection System (IDSs) provide administrators a defensive edge over intruders lodging such malicious attacks. However, with the sheer number of different IDSs available, one has to objectively assess the capabilities of different IDSs to select an IDS that meets specific organizational requirements. A prerequisite to enable such an objective assessment is the implicit comparability of IDS literature. In this study, we review IDS literature to understand the implicit comparability of IDS literature from the perspective of metrics used in the empirical evaluation of the IDS. We identified 22 metrics commonly used in the empirical evaluation of IDS and constructed search terms to retrieve papers that mention the metric. We manually reviewed a sample of 495 papers and found 159 of them to be relevant. We then estimated the number of relevant papers in the entire set of papers retrieved from IEEE. We found that, in the evaluation of IDSs, multiple different metrics are used and the trade-off between metrics is rarely considered. In a retrospective analysis of the IDS literature, we found the the evaluation criteria has been improving over time, albeit marginally. The inconsistencies in the use of evaluation metrics may not enable direct comparison of one IDS to another.

1 Introduction

Intrusion Detection Systems (IDS) have been a popular defensive strategy used in protecting entire enterprise networks, hosts in a network, or processes on a host from cyber attacks. IDSs are particularly popular because of their ability to detect insider abuse and/or outsider penetration. Over the years, the number of approaches to intrusion detection has seen a dramatic increase with over 400 literary works published in the year 2015 alone.[1]

The sheer number of intrusion detection literature provides organizations with an opportunity to choose an IDS that best suits their unique set of requirements. However, to objectively pick one intrusion detection approach over another, there is a need for the approaches to have been evaluated1. considering potential trade-off between the evaluation metrics used and 2. using evaluation metrics that may be directly compared. In this study, we systematically review intrusion detection literature to understand the extent to which the literature is comparable. We note that our study does not focus on the intrusion detection approach itself but the metrics used in its evaluation.

Research on intrusion detection systems tends not to be generalizable. Though they are quite widely used and serve an important purpose, the lack of common metrics makes it difficult to compare individual IDS's. In this study, we performed a comprehensive literature review of academic papers written about IDSs and studied the use of specific metrics that were referenced.

[1] Source: Google Scholar with title-only search for "intrusion detection system" restricted to works published in the year 2015.

The research contribution of this work is an empirical study of the intrusion detection literature. The results of this work will help guide researchers into conducting better studies, and will help guide practitioners in making a more informed decision about the intrusion detection system they analyze.

2 Research Questions

We address the following research questions in this study:

RQ0 Usage What are the commonly used evaluation metrics?

RQ1 Comparable Are empirical studies of intrusion detection systems comparable using consistent evaluation metrics?

RQ2 Trade-off Do empirical studies of intrusion detection systems consistently convey trade-off between evaluation metrics?

RQ3 Evolution Have empirical studies of intrusion detection systems improved evaluation criteria over time?

3 Related Work

In 2002, in a survey of intrusion detection research, Lundin and Jonsson [1] have highlighted the lack of methods for testing, evaluating, and comparing IDSs as one of the open research areas. The off-line intrusion detection evaluation program initiated by the Defense Advanced Research Projects Agency (DARPA) provides researchers with a testbed for consistently evaluating intrusion detection systems, however, the evaluation testbed is one factor in consistently evaluating IDSs. The testbed was used in the evaluation of several IDSs the results of which were published [2, 3]. Haines et al. [4] extended the DARPA testbed by proposing the Lincoln Adaptable Real-time Information Assurance Testbed (LARIAT). While a standardized testbed provides a common platform for evaluating IDSs, researchers are still required to execute the various IDSs to compare their performance. In our study, we are surveying the literature to understand the extent to which intrusion detection literature is implicitly comparable.

At their core, IDSs work on the principle of labeling an action as an anomaly (in case of anomaly-based detection) or misuse (in case of misuse-based detection). IDSs typically monitor some form of information stream (e.g. gateway in network-based IDS, system calls to kernel in host-based IDS). A number of researchers [5, 6, 7, 8, 9, 10] have surveyed the approaches used to label an action as an anomaly or misuse. Modi et al. [9] highlight some of the challenges of detecting intrusions in a cloud computing environment. Similarly, Anantvalee and Wu [8] and Butun et al. [10] discuss the nuances of deploying IDSs in the context of Mobile Ad Hoc Networks (MANET) and Wireless Sensor Networks (WSN), respectively. The survey of IDSs mentioned so far, while comprehensive in describing the detection approach and architecture, do not highlight the evaluation aspect of IDSs. In our study, we wish to highlight the inconsistencies that may exist in the evaluation of IDSs.

4 Methodology

As mentioned earlier, the primary focus of our study is on the metrics used in the evaluation of IDSs. At a high-level, the approach to select the primary studies for our research may be summarized as follows:

- Identify source(s) of papers
- Identify evaluation metrics
- Identify search terms for the evaluation metrics
- Collect papers using the search terms from the source(s) identified
- Filter papers based on inclusion criteria

The methodology is inspired by guidelines prescribed by Kitchenham and Charters [11].

4.1 Identify source(s) of papers

A fundamental step in any systematic literature review study is the identification of the academic sources from which the papers to review will be chosen. The most commonly used sources are the Institute of Electrical and Electronics Engineers (IEEE) and Association for Computing Machinery (ACM) publication databases. Searching for papers in IEEE and ACM databases is made possible by sophisticated web-based interfaces provided by the organizations that index the database. In our study, we have used the IEEE database as the source of papers. The IEEE Xplore (`http://ieeexplore.ieee.org/`) is the search service provided by IEEE and some of its features (like the bulk export of search results) are conducive to systematic literature reviews.

4.2 Identify Evaluation Metrics

The prerequisite to the selection of primary studies was to identify, from prior literature, the metrics commonly used in the evaluation of IDSs. As mentioned earlier, an IDS essentially *classifies* an action as being an intrusion or otherwise. The effectiveness of the classification is one among many dimensions along which an IDS may be evaluated. Therefore, in identifying the evaluation metrics, we did not constrain ourselves to particular dimension. We used IEEE Xplore with the blanket query—"intrusion detection system"—to search for all papers that mention IDS. We selected the top 1,000 papers from the IEEE Xplore search results, ordered by relevance. Three authors collectively examined 158 papers and identified the evaluation metric(s) used (if any). In the context of this step, an evaluation metric was defined as in Definition 1.

Definition 1 *Evaluation Metric—An evaluation metric is a repeatable system for expressing the quantitative evidence when reasoning about the empirical characteristics (such as effectiveness, performance, and energy consumption) of an IDS.*

Precision and *recall* are typical examples of an evaluation metric used in expressing the empirical effectiveness of an IDS. We chose to ignore metrics that were specific to an implementation. For instance, "number of generations" is a metric that is often used when genetic programming is employed to generate network intrusion detection rules based on regular expressions. While such implementation-specific metric may be relevant to future IDS implementations, we regard them as "internal" metrics and consider them as not conveying any external evidence.

We identified 22 different metrics used in the evaluation of IDSs. The evaluation metrics were grouped into four categories: (a) effectiveness, (b) performance, (c) energy, and (d) deterministic finite automaton (DFA). The categorization of the metrics enable a logical grouping of literature that evaluated an IDS along different dimensions.

4.3 Identify search terms for the evaluation metrics

The evaluation metrics identified previously were used to develop search terms. Typically, a one-to-one mapping exists between the search term and an evaluation metric with the name of the metric being the search term. However, when commonly used synonyms and/or context-specific variations of the name of a metric are considered, the mapping between the metric and the search terms becomes one-to-many. For example, the metric recall has synonyms sensitivity, hit rate, and true positive rate and a context-specific variation detection rate (in the context of an IDS).

4.4 Collect papers using the search terms from the source(s) identified

The search term(s) identified are logically combined to construct a *search string* the syntax of which is prescribed by the search engine used to search for the papers in the source(s) identified (IEEE Xplore, in our case). The search strings for each of the evaluation metrics considered in our study are presented in Appendix A. The search strings conform to the syntax prescribed by IEEE Xplore. See `http://ieeexplore.ieee.org/Xplorehelp/#/searching-ieee-xplore/search-examples#command-search-examples` for more information on the IEEE Xplore search strings.

We note that, in collecting the papers, we did not restrict the publication venue as our initial exploratory searches resulted in papers published in a variety of venues.

4.5 Filter papers based on inclusion criteria

The number of search results returned by IEEE Xplore may be controlled by applying a filter criteria (automated or manual), commonly known as the inclusion criteria. In our study, we have used the following inclusion criteria:

- Paper contains peer-reviewed academic content.
- Paper is written in English.
- Paper is at least six pages in length.
- Paper is evaluating an IDS.
- Paper uses empirically evaluation.
- Paper is indeed using the evaluation metric (specific to each search string) in the evaluation of the IDS.

The primary studies that satisfy the inclusion criteria are considered relevant and will be referred to as such in the remainder of this paper. For the most part, the inclusion criteria used in our study could not be automated to filter the primary studies. For instance, the criterion that specifies the number of pages in primary studies could be automated, albeit, to a limited extent. The limitation is due to the format of the page number metrics—starting page and ending page—obtained from IEEE Xplore. "V1-470" and "V1-474", "5 pp." and "", and "I" and "619-22 vol.1" are some of the examples of staring page number and ending page number, respectively, as returned by IEEE Xplore. The subjectivity of the other criteria necessitate reading the primary study to determine validity.

The limitations in automating the application of the inclusion criteria warranted the need for a manual approach. Since the goal is to determine the number of relevant papers, one may have to manually apply the inclusion criteria to thousands of primary studies. However, reading all the primary studies may have not been feasible. As a result, we used an approach to *estimate* the number of relevant papers using a sample of the primary studies. We used a simple sampling method to1. randomly sample a subset of the primary studies, 2. determine the proportion of relevant papers in the sample (termed *adjustment factor* in our study) by manually applying the inclusion criteria and, 3. estimate the proportion of relevant papers in the population using the sample proportion. We independently applied the approach to primary studies associated with each of the evaluation metrics considered in our study.

Our approach to determining the adjustment factor, being manual, is inherently subjective. To control the bias that may have been induced by the subjectivity, two authors independently applied the inclusion criteria to the same random sample of primary studies associated with each evaluation metric considered in our study. The authors' response to each primary study in the random sample was "include" or "exclude". We used an inter-rater reliability measure—Cohen's κ—to quantify the degree of agreement between the authors. The strength of agreement between the authors was interpreted from Cohen's κ values using the heuristic presented in a work by Landis and Koch [12]. According to the heuristic, the strength of agreement is (a) almost perfect, if $0.81 < \kappa < 1.00$, (b) substantial, if $0.61 < \kappa < 0.80$, (c) moderate, if $0.41 < \kappa < 0.60$, (d) fair, if $0.21 < \kappa < 0.40$, (e) slight, if $0.00 < \kappa < 0.20$, and (f) poor, if $\kappa < 0.00$. In cases when the strength of agreement was less than moderate, we repeated the manual application of the inclusion criteria on another random sample of the primary studies. There were two such cases—Scalability and Saved Battery—in our study. Initially, the Cohen's κ for scalability and saved battery metrics was negative, however, upon applying the inclusion criteria again we achieved an improvement in Cohen's κ in both metrics. The next step toward estimating the adjustment factor was the resolution of disagreements (if any) between the authors who applied the inclusion criteria. All primary studies that the authors disagreed on were discussed, with the authors presenting their case for including (or excluding) the primary study. Typically, the discussion led to one author convincing the other to include (or exclude) a paper. In cases where a discussion failed to yield consensus among the two authors, a third author resolved the conflict. With the disagreements resolved, the adjustment factor for each of the evaluation metrics was computed. The final value of Cohen's κ and corresponding strength of agreement and the adjustment factor computed for each of the evaluation metrics are shown in Table 1.

Table 1: Agreement between authors when manually applying the inclusion criteria to a random sample of primary studies and the adjustment factor computed after the resolution of disagreements (if any)

Metric Group	**Evaluation Metric**	**Agreement**		**Adjustment Factor**
		κ	**Strength**	
Effectiveness	Accuracy	0.75	Substantial	0.24
	Confusion Matrix	0.90	Almost Perfect	0.38
	F-value	0.83	Almost Perfect	0.13
	False Negative Rate (FNR)	0.56	Moderate	0.55
	False Positive Rate (FPR)	0.52	Moderate	0.62
	Precision	0.84	Almost Perfect	0.35
	Recall	0.88	Almost Perfect	0.33
	ROC	1.00	Almost Perfect	0.57
	True Negative Rate (TNR)	0.63	Substantial	0.38
Performance	Computational Time	0.51	Moderate	0.25
	CPU Usage	0.77	Substantial	0.23
	Memory Usage	0.88	Almost Perfect	0.09
	Scanning Time	0.59	Moderate	0.35
	Slowdown	0.50	Moderate	0.22
	Throughput	0.68	Substantial	0.27
Energy	Energy Consumption	1.00	Almost Perfect	0.16
DFA	Number of DFA States	0.50	Moderate	0.45
	Number of DFA Transitions	0.67	Substantial	0.30

We note that we removed four metrics—scalability, saved battery, expansion factor, and number of distinct DFA transitions—from our study as the number of relevant papers estimated for these metrics was negligible.

5 Results

In the subsections that follow, we address each of our research questions using the relevant papers estimated by using the adjustment factor computed by applying our methodology from Section 4.

RQ0 Usage: What are the commonly used evaluation metrics?

In this research question, we wanted to enumerate the metrics that are commonly used in the evaluation of IDS. We identified 22 different evaluation metrics, categorized into four metric groups. The definition of each of the 18 evaluation metrics and their respective metric group, as used in our study, is presented below. The definition of each of the metrics is either common knowledge or obtained from prior IDS literature. Nonetheless, each metric definition is complemented with an example of a prior IDS paper that defines and/or uses the metric.

Effectiveness Metrics

The metrics that belong to this category express the effectiveness of an IDS in labeling an action as an anomaly, misuse, or intrusion. The metrics in this category are defined in terms of the types of correct or erroneous classifications that an IDS can make. The outcome from IDS, being a binary classifier, can be one of (a) true positive (TP), if an intrusive action is classified as such, (b) true negative (TN), if a legitimate action is classified as such, (c) false positive (FP), if a legitimate action is classified as intrusive, or (d) false negative (FN), if an intrusive action is classified as legitimate. We have nine metrics that belong to this category, they are:

1. *Confusion Matrix* is a 2×2 matrix that contains the values of TP, TN, FP, and TN. A typical confusion matrix takes the form show in Table 2. A confusion matrix can also be $n \times n$ in dimension in the case of a multi-class classifier with n different classes. For instance, in applying a multi-class machine learning model to detecting 20 different classes of malicious web activities, Goseva-Popstojanova et al. [13] have presented a 20×20 confusion matrix to assert the effectiveness of the model.

   Table 2: Typical confusion matrix

		Prediction	
		Intrusion	**Legitimate**
Truth	**Intrusion**	TP	FN
	Legitimate	FP	TN

   While the confusion matrix in itself is *not* a metric, it is a container of metrics that may be used to compute other effectiveness metrics.

2. *Accuracy* is the proportion of correct classifications (includes both TP and TN). Accuracy may be computed using (1) with the values of TP, FP, TN, and FN obtained from the confusion matrix. Accuracy is used as an evaluation metric in a work by Ali and Al-Shaer [14].

$$accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{1}$$

3. *Precision* is the proportion of classified actions that are actually intrusive. Precision may be computed using (2) with the values of TP and FP obtained from the confusion matrix. Precision is used as an evaluation metric in a work by Tang et al. [15].

$$precision = \frac{TP}{(TP + FP)} \tag{2}$$

4. *Recall* is the proportion of intrusive actions that are classified as intrusive. Recall may be computed using (3) with the values of TP and FN obtained from the confusion matrix. Recall is used as an evaluation metric in a work by Tang et al. [15].

$$recall = \frac{TP}{(TP + FN)} \tag{3}$$

5. *F-value* is the harmonic mean of precision and recall. F-value may be computed using (4) with precision and recall computed using (2) and (3), respectively. F-value is used as an evaluation metric in a work by Gupta et al. [16].

$$f\text{-}value = 2 \cdot \frac{precision \cdot recall}{(precision + recall)} \tag{4}$$

6. *False Positive Rate (FPR)* may be computed using (5) with the values of FN and TP obtained from the confusion matrix. FPR is used as an evaluation metric in a work by Bartoli et al. [17].

$$fpr = \frac{FP}{(FP + TN)} \tag{5}$$

7. *False Negative Rate (FNR)* may be computed using (6) with the values of FN and TP obtained from the confusion matrix. FNR is used as an evaluation metric in a work by Bartoli et al. [17].

$$fnr = \frac{FN}{(FN + TP)} \tag{6}$$

8. *True Negative Rate (TNR)* may be computed using (7) with the values of TN and FP obtained from the confusion matrix. TNR is used as an evaluation metric in a work by Sharma et al. [18].

$$tnr = \frac{TN}{(TN + FP)} \tag{7}$$

9. *Receiver Operator Characteristic (ROC)* is a plot of recall as a function of fall-out (i.e. proportion of legitimate actions classified as such). ROC curve, as the plot is called, is a powerful tool to evaluate the effectiveness of an IDS while considering the trade-off between recall and fall-out. ROC is used as an evaluation tool in a work by Salem et al. [19].

Performance Metrics

The metrics that belong to this category express the empirical performance of an IDS in the context of a testbed. We have seven metrics that belong to this category, they are:

1. *Computational time* is the time taken to complete an essential task in classifying an action as intrusion or legitimate. In evaluating a clustering-based IDS, Owezarski [20] expressed the time taken to determine clusters as the computational time.

2. *CPU usage* is the percentage of load on the CPU with the addition of an IDS to the infrastructure. In evaluating an sensor embedded IDS, Maciá-Pérez et al. [21] measured the load on the CPU during various network bandwidths to assert the performance of their IDS.

3. *Memory usage* is the amount of memory required by an IDS to perform its classification. In evaluating an novel approach to string matching in the presence of out-of-sequence pattern matching, Chen et al. [22] compared the memory usage (in Mega Bytes) of their approach with a state-of-the-art approach.

4. *Scanning time* is the time spent scanning patterns among pattern rulesets looking for a matching pattern. In evaluating an approach to high-speed pattern matching in the context of an IDS, Choi and Seo [23] have used scanning time as a metric.

5. *Slowdown* is the additional time that a network/host/process takes to complete an operation while being monitored by an IDS. In evaluating an approach to detecting malicious modifications to operating system data by kernel rootkits, Hofmann et al. [24] used slowdown as a metric to express the impact such an approach may have on the operation of the system.

6. *Throughput* is the amount of data that can be processed by an IDS per second. In evaluating a memory-efficient pattern matching algorithm, Dharmapurikar and Lockwood [25] used throughput (in Giga Bits per Second) to assert the performance improvement exhibited by their algorithm.

Energy Metrics

The metrics that belong to this category express the energy consumption of an IDS. The metrics in this category are typically used in the context of an IDS implementation in Mobile Ad-hoc Network (MANET), Wireless Sensor Network (WSN), and Vehicular Ad-hoc Network (VANET). We have two metrics that belong to this category, they are:

1. *Energy consumption* is the additional energy consumed by a device, typically mobile, with the introduction of an IDS. In evaluating a decentralized IDS in a wireless sensor network, da Silva et al. [26] measured the increase in energy consumption by a monitor node before and after the activation of an IDS component.

Deterministic Finite Automata (DFA) Metrics

The metrics that belong to this category express the aspects related to the implementation of an IDS using Deterministic Finite Automata (DFA). We have four metrics that belong to this category, they are:

1. *Number of DFA states* is a metric that quantifies the size of a DFA. Number of DFA states is used as an evaluation metric in a work by Kumar et al. [27].

2. *Number of DFA transitions* is a metric that quantifies the different paths of transition between states in a DFA. Number of DFA transitions is used as an evaluation metric in a work by Kumar et al. [27].

RQ1 Comparable: Are empirical studies of intrusion detection systems comparable using consistent evaluation metrics?

In this research question, we wanted to understand the metrics that were most commonly used in the evaluation of IDSs. We address this question by comparing the number of relevant papers that are associated with a particular evaluation metric. To estimate the number of relevant papers, we simply multiply the number of primary studies associated with each evaluation metric and the corresponding adjustment factor presented in Table 1. The number of primary studies and the number of relevant papers in each of the 18 evaluation metrics is presented in Table 3 in Appendix B. Shown in Figure 1 are all the 18 evaluation metrics, grouped by metric category, ordered in descending order of the number of relevant papers that use the metric in the evaluation of an IDS.

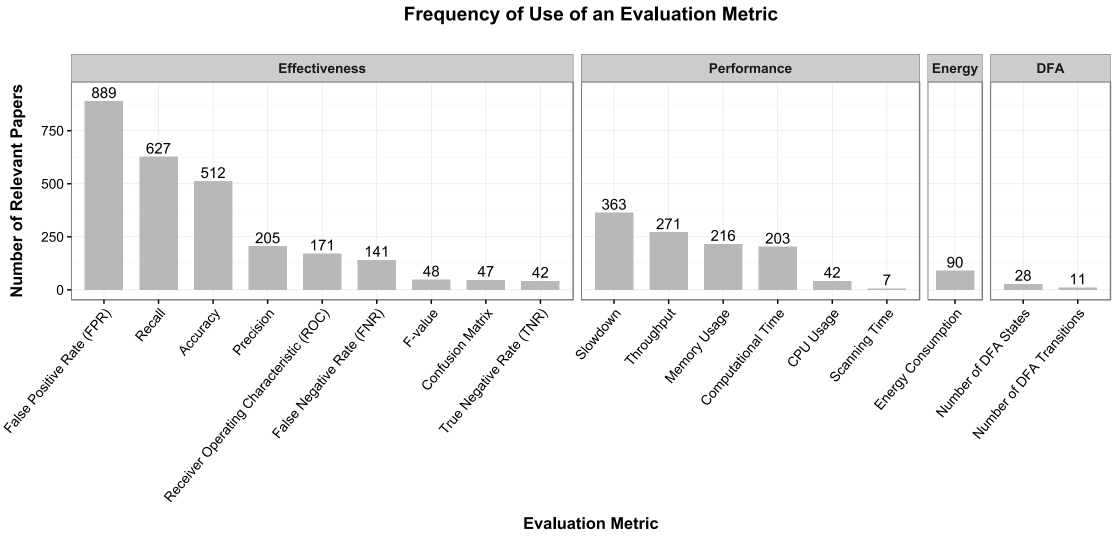

Figure 1: Evaluation metrics ordered by the frequency of usage

As seen in the figure, false positive rate and recall effectiveness metrics, slowdown and throughput performance metrics, energy consumption metric, and number of DFA states DFA metric are the most commonly used evaluation metrics. The popularity of certain metrics indicate the most common concerns in evaluating an IDS. For instance, having a low false positive rate is a critical to ensuring that the IDS itself does not become a source of denial of service to legitimate actions. While few metrics are commonly used, there still are a high number of other metrics that have been used in considerably large number of relevant papers. The spectrum of metrics does not facilitate a direct comparison between different relevant papers preventing an objective evaluation of IDSs for implementation by an organization.

RQ2 Trade-offs: Do empirical studies of intrusion detection systems consistently convey trade-offs between evaluation metrics?

In this research question, we wanted to understand the extent to which intrusion detection literature considered, and reported, trade-offs between evaluation metrics. The trade-off between evaluation metrics is an important consideration when one is evaluating an IDS. Consider the trade-off between the commonly used effectiveness metrics: precision and recall. As mentioned earlier, in the context of an IDS, precision is the proportion of intrusive actions identified that were actually intrusions and recall quantifies the proportion of all intrusive actions that were identified. An IDS could achieve a precision of 100% if at least one action is correctly marked as intrusive. Similarly, an IDS could achieve a recall of 100% if all actions are marked as intrusive. Clearly, an IDS with a precision of 100% or a recall of 100% does not guarantee that all intrusive actions are marked as such.

We address this research question using venn diagrams to present the number of papers that use two or more metrics know to trade-off one another. In RQ1, we estimated the number of relevant papers using the adjustment factor. Here again, we use a similar approach to estimate the number of relevant papers that use multiple evaluation metrics but with an additional step: the adjustment factors presented in Table 1 were computed for primary studies associated with a single evaluation metric. Since the adjustment factor is analogous to the probability of finding a relevant paper in a random sample of primary studies, we multiplied the adjustment factors associated with two or more metrics to compute the adjustment factor that represents the probability of finding a relevant paper that uses the metrics in question. For instance, the adjustment factors for precision and recall are 0.35 and 0.33, respectively. The adjustment factor $0.1155 = 0.35 \cdot 0.33$ may be used to estimate the number of relevant papers that use both precision and recall.

Figure 2 shows the trade-off between precision and recall. As seen in the figure, a small number of relevant papers were estimated to be using both precision and recall.

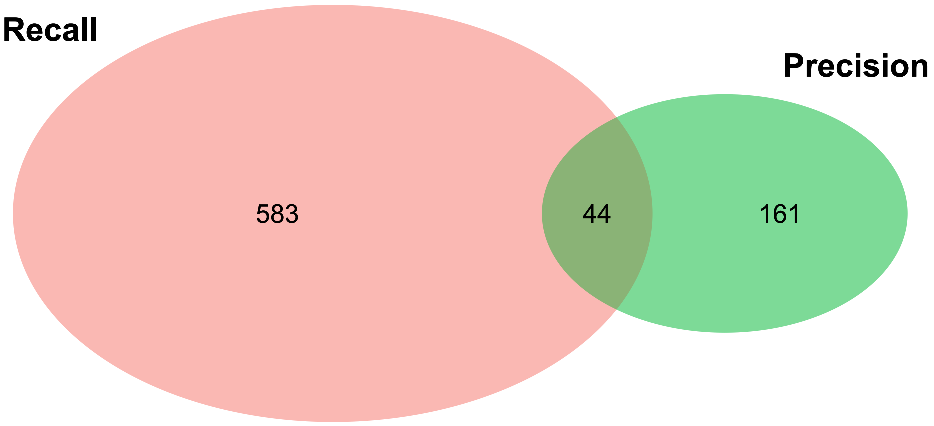

Figure 2: Trade-off between precision and recall

The F-value, being the harmonic mean of precision and recall, may be used to convey the trade-off between precision and recall. However, presenting all three metrics may enable better interpretation of the effectiveness of the IDS. From the Figure 3, we see that only a small number of papers that present precision and recall also present F-value.

In the context of performance, the trade-off between space (memory usage) and time (computational time) is the one that is most commonly discussed. In evaluating an IDS from a performance perspective, reporting the improvement in computational time without discussing the potential increase in memory usage

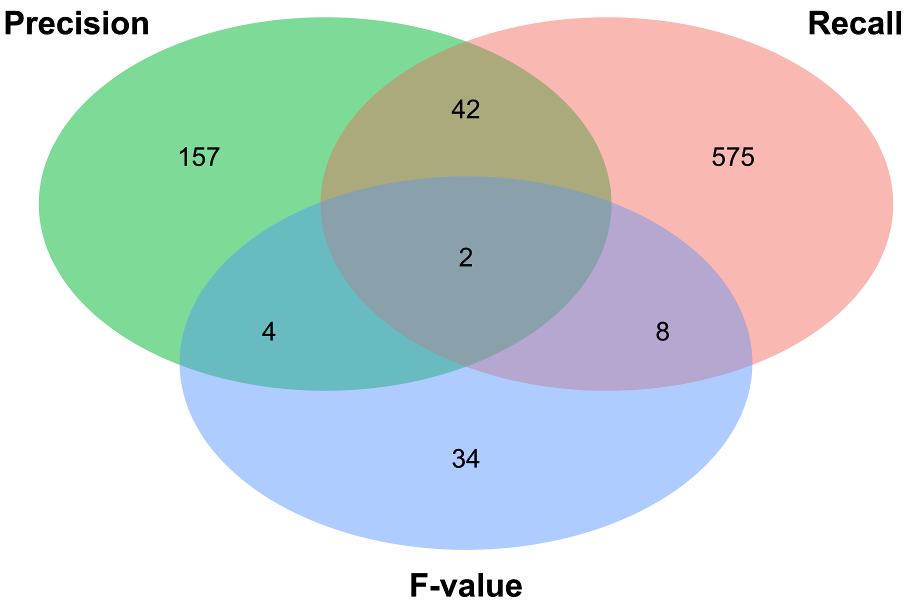

Figure 3: Trade-off between precision, recall, and F-value

does not enable the objective interpretation of the IDS' performance. Shown in Figure 4 is the number of papers that consider this trade-off and report both computational time and memory usage.

The figures reveal that little consideration is given to evaluation metrics that tend to trade-off one another. As a result, the evaluation may be biased toward a certain aspect.

RQ3 Evolution: Have empirical studies of intrusion detection systems improved evaluation criteria over time?

In RQ2, we noted the lack of consideration given to trade-off between metrics when evaluation IDSs. In this research question, we wanted to understand if the evaluation criteria have been improving over the years, getting better at considering trade-off between metrics. In other words, we wanted to understand if the number of relevant papers using two or more metrics that tend to trade-off has been increasing over time relative to the number of relevant papers using the metrics independently.

In the context of this research question, we applied the adjustment factor to the number of primary studies in each year. The analysis started with the first year in which all metrics considered had at least one primary study (For example, we had one primary study using recall published in the year 1944, however, there were no papers that used precision published until the year 1982. Therefore, we used 1982 as the first year in our analysis).

Shown in Figure 5 is an area plot showing the number of relevant papers that used precision, recall, and both precision and recall. While the number of relevant papers is low between years 1982 and 1999, all the papers published in these years used recall as the evaluation metric. We also notice that the number of relevant papers using both precision and recall is increasing, albeit marginally.

We did perform similar analysis for the performance metrics—computational time and memory usage—but the number of relevant papers in each year was zero for most years to draw any conclusions.

6 Threats to Validity

In this section, we present some of the threats to validity of our study and the way in which we have mitigated the threats.

Single Source of Primary Studies

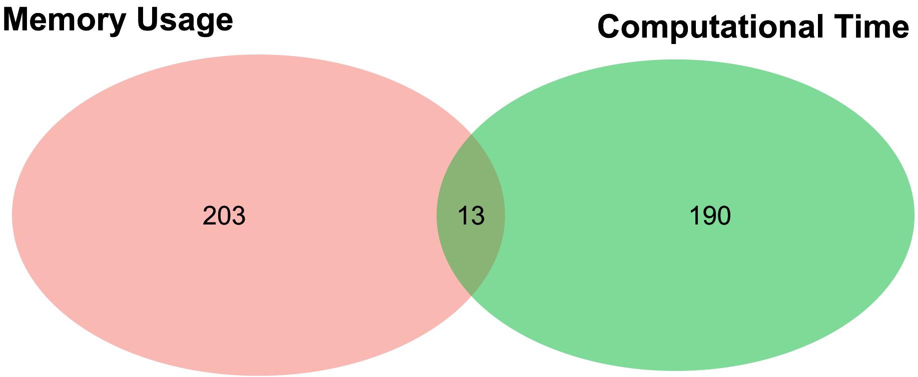

Figure 4: Trade-off between computational time and memory usage

Although the guidelines [28] recommends the usage of at least two sources (ACM Digital Library and IEEE Xplore) of academic content, the lack of a convenient way of exporting search results from ACM Digital Library prevented us from using it as a source of primary studies. An attempt at scraping the search results of the website led to our IP address being blocked for several days and an email to ACM did not yield a response.

Subjectivity

Subjectivity is inherent in any manual approach, however, we can control the bias induced by such inherent subjectivity. We have used inter-rater reliability measure—Cohen's κ—to quantify, and control, the bias in scenarios where a manual approach was used.

False Positives in Search Results

A simple full-text search for the phrase "intrusion detection systems" on IEEE Xplore results in more than 7,000 papers being returned. However, a large proportion of these papers may not be evaluating an IDS and, therefore, be regarded as false positives in the context of our study. Several papers (For example, the work by Strunk et al. [29]) merely mention intrusion detection systems in passing or as a possible application of their algorithm. In our study, we have manually analyzed a random sample of the primary studies returned by IEEE Xplore to determine a proportion of those that are relevant in the context of IDSs.

Publication Quality

We acknowledge that not every paper published by IEEE are reviewed with the same peer-review standards. We did not, however, consider the *quality* of the publications as there are only proxies (such as number of citations per year and acceptance rate) that have been used to quantify the quality. We have used a common but crude approach to include papers that contain sufficient detail by removing papers that have fewer than six pages.

Correctness of Analysis Scripts and Data Aggregation

The search results from IEEE Xplore were exported to CSV files and analyzed using a Python application (Available as an open-source project at `https://github.com/andymeneely/CrossSearchCrawler`). We conducted code reviews throughout the development of the application and also tested the code manually to ensure correctness.

IEEE Xplore is sensitive to seemingly innocuous changes; including an additional parenthesis could result in the inclusion or exclusion of hundreds of results. We mitigated this limitation by ensuring that the search

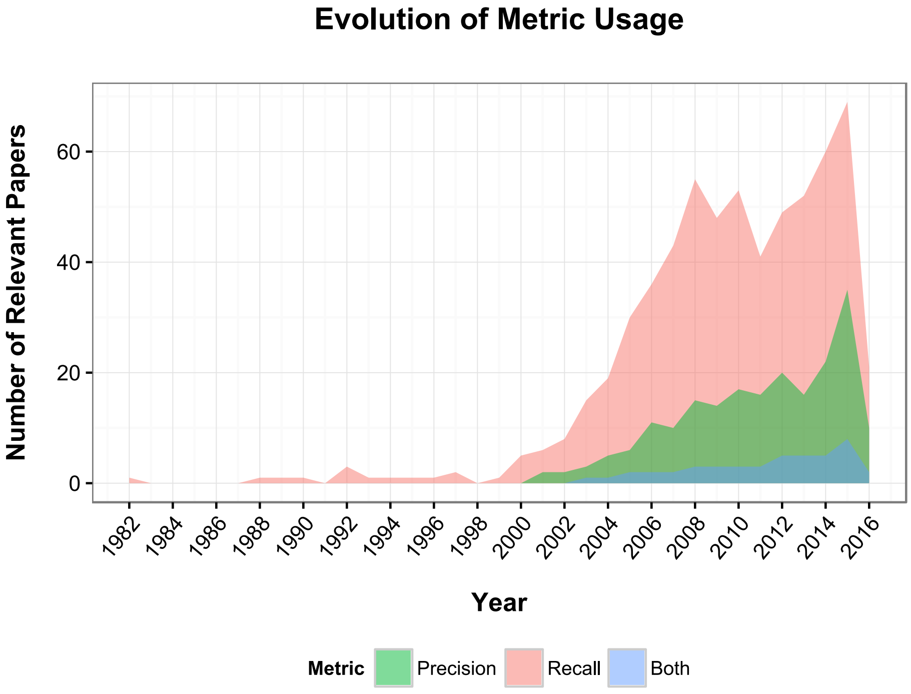

Figure 5: The evolution of consideration to trade-off between precision and recall

strings adhered to the syntax prescribed by IEEE Xplore.

7 Summary

In this systematic literature review of intrusion detection literature, we aimed to explore the use of metrics in the evaluation of IDSs, specifically the consistency of metric usage. We manually reviewed a set of 158 papers and identified 22 metrics that are commonly used in the evaluation of IDSs. We constructed search strings for each of the 22 metrics and using IEEE Xplore to search for primary studies. We manually reviewed 495 primary studies returned by IEEE Xplore and found 159 of them to be relevant. We used the manual review to estimate the total number of relevant papers in all primary studies returned by IEEE Xplore. We then analyzed the number of relevant papers in addressing four research questions.

The results of our review show the inconsistencies that exist in the evaluation of IDS, specifically in the context of the metrics used. We found that multiple different metrics are used in the evaluation of IDSs (RQ1) and oftentimes the trade-off between the metrics is not considered (RQ2). A retrospective analysis of the inconsistencies also revealed that, while the evaluation criteria has been improving over time, the improvement is marginal at best.

8 Acknowledgment

This research was sponsored by the U.S. National Security Agency Science of Security Lablet at North Carolina State University. The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of any sponsoring institution, the U.S. government or any other entity.

References

[1] E. Lundin and E. Jonsson, "Survey of Intrusion Detection Research," Chalmers University of Technology, Tech. Rep., 2002.

[2] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. R. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham, and Others, "Evaluating Intrusion Detection Systems: The 1998 DARPA Off-line Intrusion Detection Evaluation," in *DARPA Information Survivability Conference and Exposition, 2000. DISCEX'00. Proceedings*, vol. 2. IEEE, 2000, pp. 12–26.

[3] R. Lippmann, J. W. Haines, D. J. Fried, J. Korba, and K. Das, "The 1999 DARPA Off-Line Intrusion Detection Evaluation," *Computer networks*, vol. 34, no. 4, pp. 579–595, 2000.

[4] J. W. Haines, L. M. Rossey, R. P. Lippmann, and R. K. Cunningham, "Extending the DARPA Off-line Intrusion Detection Evaluations," in *DARPA Information Survivability Conference & Exposition II, 2001. DISCEX'01. Proceedings*, vol. 1. IEEE, 2001, pp. 35–45.

[5] T. F. Lunt, "A survey of intrusion detection techniques," *Computers & Security*, vol. 12, no. 4, pp. 405–418, 1993.

[6] S. Axelsson, "Intrusion Detection Systems: A Survey and Taxonomy," Tech. Rep., 2000.

[7] A. Murali and M. Rao, "A Survey on Intrusion Detection Approaches," in *2005 International Conference on Information and Communication Technologies*. IEEE, 2005, pp. 233–240.

[8] T. Anantvalee and J. Wu, "A Survey on Intrusion Detection in Mobile Ad Hoc Networks," in *Wireless Network Security*. Springer, 2007, pp. 159–180.

[9] C. Modi, D. Patel, B. Borisaniya, H. Patel, A. Patel, and M. Rajarajan, "A survey of intrusion detection techniques in cloud," *Journal of Network and Computer Applications*, vol. 36, no. 1, pp. 42–57, 2013.

[10] I. Butun, S. D. Morgera, and R. Sankar, "A Survey of Intrusion Detection Systems in Wireless Sensor Networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 266–282, 2014.

[11] B. Kitchenham and S. Charters, "Guidelines for performing Systematic Literature Reviews in Software Engineering," in *Evidence Based Software Engineering Technical Report*, 2007.

[12] J. R. Landis and G. G. Koch, "The Measurement of Observer Agreement for Categorical Data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977. [Online]. Available: http://www.jstor.org/stable/2529310

[13] K. Goseva-Popstojanova, G. Anastasovski, and R. Pantev, "Using Multiclass Machine Learning Methods to Classify Malicious Behaviors Aimed at Web Systems," in *2012 IEEE 23rd International Symposium on Software Reliability Engineering*, Nov 2012, pp. 81–90.

[14] M. Q. Ali and E. Al-Shaer, "Configuration-based IDS for Advanced Metering Infrastructure," in *Proceedings of the 2013 ACM SIGSAC Conference on Computer &#38; Communications Security*, ser. CCS '13. New York, NY, USA: ACM, 2013, pp. 451–462. [Online]. Available: http://doi.acm.org/10.1145/2508859.2516745

[15] L.-A. Tang, X. Yu, Q. Gu, J. Han, G. Jiang, A. Leung, and T. L. Porta, "A Framework of Mining Trajectories from Untrustworthy Data in Cyber-Physical System," *ACM Trans. Knowl. Discov. Data*, vol. 9, no. 3, pp. 16:1—-16:35, feb 2015. [Online]. Available: http://doi.acm.org/10.1145/2700394

[16] K. K. Gupta, B. Nath, and R. Kotagiri, "Layered Approach Using Conditional Random Fields for Intrusion Detection," *IEEE Transactions on Dependable and Secure Computing*, vol. 7, no. 1, pp. 35–49, Jan 2010.

[17] A. Bartoli, G. Davanzo, and E. Medvet, "A Framework for Large-Scale Detection of Web Site Defacements," *ACM Trans. Internet Technol.*, vol. 10, no. 3, pp. 10:1–10:37, Oct. 2010. [Online]. Available: http://doi.acm.org/10.1145/1852096.1852098

[18] R. Sharma, P. Sharma, P. Mishra, and E. S. Pilli, "Towards MapReduce based classification approaches for Intrusion Detection," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Jan 2016, pp. 361–367.

[19] O. Salem, A. Makke, J. Tajer, and A. Mehaoua, "Flooding attacks detection in traffic of backbone networks," in *Local Computer Networks (LCN), 2011 IEEE 36th Conference on*, Oct 2011, pp. 441–449.

[20] P. Owezarski, "A Near Real-Time Algorithm for Autonomous Identification and Characterization of Honeypot Attacks," in *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, ser. ASIA CCS '15. New York, NY, USA: ACM, 2015, pp. 531–542. [Online]. Available: http://doi.acm.org/10.1145/2714576.2714580

[21] F. Maciá-Pérez, F. J. Mora-Gimeno, D. Marcos-Jorquera, J. A. Gil-Martinez-Abarca, H. Ramos-Morillo, and I. Lorenzo-Fonseca, "Network Intrusion Detection System Embedded on a Smart Sensor," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 3, pp. 722–732, mar 2011.

[22] X. Chen, K. Ge, Z. Chen, and J. Li, "AC-Suffix-Tree: Buffer Free String Matching on Out-of-Sequence Packets," in *Proceedings of the 2011 ACM/IEEE Seventh Symposium on Architectures for Networking and Communications Systems*, ser. ANCS '11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 36–44. [Online]. Available: http://dx.doi.org/10.1109/ANCS.2011.14

[23] Y. H. Choi and S. W. Seo, "BLAST: B-LAyered bad-character SHIFT tables for high-speed pattern matching," *IET Information Security*, vol. 7, no. 3, pp. 195–202, Sept 2013.

[24] O. S. Hofmann, A. M. Dunn, S. Kim, I. Roy, and E. Witchel, "Ensuring Operating System Kernel Integrity with OSck," in *Proceedings of the Sixteenth International Conference on Architectural Support for Programming Languages and Operating Systems*, ser. ASPLOS XVI. New York, NY, USA: ACM, 2011, pp. 279–290. [Online]. Available: http://doi.acm.org/10.1145/1950365.1950398

[25] S. Dharmapurikar and J. W. Lockwood, "Fast and Scalable Pattern Matching for Network Intrusion Detection Systems," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 10, pp. 1781–1792, Oct 2006.

[26] A. P. R. da Silva, M. H. T. Martins, B. P. S. Rocha, A. A. F. Loureiro, L. B. Ruiz, and H. C. Wong, "Decentralized Intrusion Detection in Wireless Sensor Networks," in *Proceedings of the 1st ACM International Workshop on Quality of Service &Amp; Security in Wireless and Mobile Networks*, ser. Q2SWinet '05. New York, NY, USA: ACM, 2005, pp. 16–23. [Online]. Available: http://doi.acm.org/10.1145/1089761.1089765

[27] S. Kumar, J. Turner, and J. Williams, "Advanced Algorithms for Fast and Scalable Deep Packet Inspection," in *Architecture for Networking and Communications systems, 2006. ANCS 2006. ACM/IEEE Symposium on*, Dec 2006, pp. 81–92.

[28] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Information and Software Technology*, vol. 55, no. 12, pp. 2049–2075, 2013. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2013.07.010

[29] J. D. Strunk, G. R. Goodson, M. L. Scheinholtz, C. A. N. Soules, and G. R. Ganger, "Self-Securing Storage: Protecting Data in Compromised Systems," in *Foundations of Intrusion Tolerant Systems, 2003 [Organically Assured and Survivable Information Systems]*, 2003, pp. 195–209.

A Search Strings

Accuracy accuracy AND “intrusion detection system”

Confusion Matrix “confusion matrix” AND “intrusion detection system”

F-value (“f-value” OR “f-measure” OR “f-score” OR “f1” OR “f2” OR “f0.5” OR “detection effectiveness”) AND “intrusion detection system”

False Negative Rate (FNR) (“false negative rate” OR “type ii error” OR “miss rate”) AND “intrusion detection system”

False Positive Rate (FPR) (“false positive rate” OR “false positive frequency” OR “false alarm” OR “type i error” OR “fall-out”) AND “intrusion detection system”

Precision (“precision” OR “positive predictive value”) AND “intrusion detection system”

Recall (“recall” OR “sensitivity” OR “true positive rate” OR “detection rate” OR “hit rate”) AND “intrusion detection system”

Receiver Operating Characteristic (ROC) (“roc” OR “receiver operating characteristic” OR “area under curve” OR “area under the curve” OR “area under roc” OR “area under the roc” OR “auc” OR “auroc”) AND “intrusion detection system”

True Negative Rate (TNR) (“true negative rate” OR “correct rejection” OR “specificity”) AND “intrusion detection system”

Computational Time (“computational time” OR “execution time” OR “running time” OR “processing time”) AND “intrusion detection system”

CPU Usage (“cpu usage” OR “processor usage”) AND “intrusion detection system”

Memory Usage (“memory usage” OR “memory”) AND “intrusion detection system”

Scalability (“number of resources” OR “number of nodes”) AND “intrusion detection system”

Scanning Time (“scanning time” OR “time spent scanning” OR “time spent pattern matching” OR “pattern matching time”) AND “intrusion detection system”

Slowdown (“slowdown” OR “overhead” OR “change in system speed” OR “increase in processing time”) AND “intrusion detection system”

Throughput (“throughput” OR “processing speed” OR “processing bandwidth” OR “bytes analyzed per second”) AND “intrusion detection system”

Energy Consumption (“energy consumption” OR “power consumption”) AND “intrusion detection system”

Saved Battery (“battery lifetime” OR “battery drain”) AND “intrusion detection system”

Expansion Factor expansion factor AND “dfa” AND “intrusion detection system”

Number of Distinct DFA Transitions (“number of distinct transitions” OR “# of distinct transitions”) AND “dfa” AND “intrusion detection system”

Number of DFA States (“number of states” OR “num of states”) AND “dfa” AND “intrusion detection system”

Number of DFA Transitions (“number of transitions” OR “# of transitions”) AND “dfa” AND “intrusion detection system”

B Number of Primary Studies and Relevant Papers

Table 3: Number of primary studies returned by IEEE Xplore and the number of relevant papers estimated using the adjustment factor computed from manual review of a random sample of primary studies

Metric Group	Evaluation Metric	# Primary Studies	# Relevant Papers
Effectiveness	Accuracy	2099	512
	Confusion Matrix	123	47
	F-value	387	48
	False Negative Rate (FNR)	258	141
	False Positive Rate (FPR)	1440	889
	Precision	593	205
	Recall	1881	627
	ROC	302	171
	True Negative Rate (TNR)	111	42
Performance	Computational Time	811	203
	CPU Usage	184	42
	Memory Usage	2320	216
	Scanning Time	20	7
	Slowdown	1635	363
	Throughput	1015	271
Energy	Energy Consumption	564	90
DFA	Number of DFA States	62	28
	Number of DFA Transitions	38	11