A Taxonomy of Metrics and Tests to Evaluate and Validate Properties of Industrial Intrusion Detection Systems

Cyntia Vargas Martinez[1][a] and Birgit Vogel-Heuser[2][b]
[1]*Bosch Rexroth AG, Bgm.-Dr.-Nebel-Str. 2, 97816 Lohr am Main, Germany*
[2]*Institute of Automation and Information Systems, Technical University of Munich,*
Boltzmannstr. 15, 85748 Garching bei München, Germany
cyntia.vargasmartinez2@boschrexroth.de, vogel-heuser@tum.de

Keywords: Information Security, Industrial Automation Systems, Intrusion Detection, Network Security, System Testing, System Validation.

Abstract: The integration of Intrusion Detection Systems (IDS) in Industrial Automation Systems (IAS) has gained popularity over the past years. This has occurred due to their ability to detect intrusions at a device and network level. In order for these systems to provide effective and reliable protection, they must possess a set of specific properties. These properties are inherent characteristics that depend on the IDS application field, as different fields provide different deployment conditions. Unfortunately, the evaluation and validation of such properties for IAS has proven challenging, as current contributions often follow evaluation and validation approaches from the IT domain that focus solely on the effectiveness of intrusion detection approaches; hence, neglecting other aspects relevant to the industrial domain. This paper addresses this issue by presenting IDS properties derived from trends and characteristics of IAS; as well as a taxonomy of metrics and tests to evaluate and validate these properties. This taxonomy provides a foundation from which future IDS contributions for IAS can be improved and reinforced by providing an overview of pertinent metrics and tests.

1 INTRODUCTION

Intrusion Detection Systems (IDS) are software applications capable of monitoring and analyzing events and information from hosts and/or the network in order to detect intrusions. Some contemporary IDS contributions have claimed a high detection rate of over 95% (Tavallaee et al., 2010). Unfortunately, although this high detection rate is their biggest allurement, some other aspects have been neglected.

This is especially highlighted in (Paxson, 2007; Tavallaee et al., 2010; Sommer and Paxson, 2010; Bhuyan et al., 2014). Where the aspect most often discussed is the qualitative and quantitative evaluation of performance that has been hindered by the lack of available datasets and the suitability of the evaluation metrics. Other aspects that have been neglected are properties that any IDS should possess. Some of these properties (i.e., soundness, completeness, timeliness, etc.) were first described in earlier works by pioneers in the field of IDS as open issues that need to be addressed (Paxson, 2007; McHugh, 2001; Bhuyan et al., 2014). Although these open issues have been known for decades in the IT domain, their neglect has inherited inadequate and/or incomplete evaluation and validation practices for the industrial domain. The addressing of such open issues is imperative in industrial IDS, as they may result in vulnerable or flawed IDS being deployed resulting in embedded vulnerabilities or misidentification of intrusions in IAS.

The main contribution of this paper is a taxonomy of metrics and tests that can be used to evaluate and validate properties that industrial IDS should possess. These properties are abstracted from the aforementioned open issues and the general characteristics and features of current industrial IDS that contemplate general architectures of IAS. The metrics and tests are identified through a literature review of IDS contributions in both the IT and industrial domains.

This paper is structured as follows. Section 2 provides a brief overview of the current state of intrusion detection in IAS. Section 3 describes abstracted properties that an IDS should possess based on the characteristics and requirements of IAS. Section 4 presents metrics and tests that can be used to evaluate and val-

[a] https://orcid.org/0000-0002-4809-4304
[b] https://orcid.org/0000-0003-2785-8819

Martinez, C. and Vogel-Heuser, B.
A Taxonomy of Metrics and Tests to Evaluate and Validate Properties of Industrial Intrusion Detection Systems.
DOI: 10.5220/0007833902010210
In *Proceedings of the 16th International Joint Conference on e-Business and Telecommunications (ICETE 2019)*, pages 201-210
ISBN: 978-989-758-378-0

idate the properties derived in Section 3. Section 5 presents the proposed Taxonomy of metrics and tests. Finally, Section 6 provides the conclusions.

2 OVERVIEW OF INTRUSION DETECTION IN IAS

This Section provides an overview of research contributions, as well as Open Source and Commercial products, in the field of intrusion detection in IAS.

2.1 Research Contributions

As it can be observed from survey literature (Zhu and Sastry, 2010; Garitano et al., 2011; Mitchell and Chen, 2014), three intrusion detection approaches are predominant in research works.

The first approach is the preference of network intrusion detection over host intrusion detection as observed in (Zhu and Sastry, 2010; Mitchell and Chen, 2014). A Network IDS (NIDS) often consists of a centralized software component that processes and analyzes all data collected from the network in order to detect intrusions and distributed units (often called sensors) that capture (and sometimes pre-process or filter) captured network data and forward it to the centralized software component. The preference of network intrusion detection over host intrusion detection comes from these distributed units that allow for passive monitoring to be performed with the use of network taps or port mirroring in switches, which allows seamless integration in legacy systems and systems during runtime. It also provides scalability and does not add overhead that may negatively influence the automation system itself. Moreover, network data collection is performed in strategical parts of the system (e.g., between HMI and SCADA System and between HMI and PLC (Kleinmann and Wool, 2017)) in order to decrease costs and increase efficiency, as only relevant and desired information is analyzed.

The second predominant approach is the preference of anomaly-based over signature-based intrusion detection. In signature-based intrusion detection data or events collected are compared to well-known intrusion patterns. On the other hand, in anomaly-based intrusion detection intrusions are not known; hence, everything that deviates from the normal behavior is considered as an intrusion (Udd et al., 2016). Anomaly-based intrusion detection is favored as signature-based intrusion detection is ineffective against zero-day attacks and is also considered costly due to the efforts required to generate its signatures.

The third, and final, most predominant approach is the preference to model automation process variables in order to detect system anomalies assumed to be caused by intrusions. This is often achieved by extracting process variables from captured network traffic (Nivethan and Papa, 2016). The predominance of these approaches in IAS is a result of the long lifetime of this kind of systems and their often predictable behavior (Naedele and Biderbost, 2004).

Some of the most popular techniques implemented to detect network intrusions are the following. Specific signatures have been generated for industrial protocols in order to detect abnormal protocol behavior (Yang et al., 2014). Normal user and service behavior have been modeled using statistical analysis from header information in (Kwon et al., 2015) and (Valdes and Cheung, 2009) respectively. More complex analysis techniques have implemented classifiers (Zhang et al., 2016; Ponomarev and Atkison, 2016), rule generators (Udd et al., 2016; Littler et al., 2013), One-Class Support Vector Machines (OCSVM) (Maglaras and Jiang, 2014), etc.

2.2 Open Source and Commercial industrial IDS

The two most widely known and used Open Source NIDS are Zeek (formerly known as Bro) (Paxson, 1999) and Snort (Roesch, 1999). They are capable of capturing, logging and analyzing network data. They also have a predefined set of signatures to detect well-known threats. However, they can also be extended to add personalized signatures or additional behavior.

Zeek filters unnecessary network data and feeds it to an interpreter that evaluates it against scripts written in the Bro scripting language (Paxson, 1999). This allows Zeek to be extended as observed in (Udd et al., 2016) where a parser was generated for IEC 60870-5-104 protocol support. On the other hand, Snort handles rules written in the Snort format. Preprocessors or plug-ins can also be integrated on it to add additional behavior. Within the context of the Digital Bond Project Quickdraw, Snort rules for industrial protocols (i.e., DNP3, Ethernet/IP and Modbus TCP) were generated (Littler et al., 2013). These rules have also been integrated into commercial IDS (Mahan et al., 2011). Zeek and Snort also share other similarities with one-another, such as the configurability of their intrusion responses which can be to either monitor traffic and block and/or report intrusions.

Commercial products with intrusion detection capabilities are often not marketed as industrial IDS but as Centralized Management solutions with a strong Cybersecurity focus or as Operational Technology

(OT) Management Systems; as their capabilities are not only limited to intrusion detection, but they also provide additional features that allow them to manage, monitor and visualize other system components. Nonetheless, they are referred to as industrial IDS in this paper as their architecture and behavior are similar to that of the research and open source NIDS previously discussed (i.e., they consist of Network Sensors and a centralized software component).

Other capabilities commonly provided by these solutions are logging, alerts, reports and graph generation that allow to visualize system events (e.g., intrusions detected). Some solutions provide more complex capabilities such as interoperability with other devices and applications such as firewalls and other Security Information and Event Management Systems (SIEM) and vulnerability assessment tools. A list of some fo these solutions can be found in the Market Guide for Operational Technology Security 2017 (Perkins et al., 2017) by Gartner, Inc.

Unfortunately, more technical information regarding the implemented techniques, operations and computational performance of these solutions is not publicly available and hence difficult to obtain.

3 INDUSTRIAL IDS PROPERTIES

This section introduces properties that industrial IDS should possess. These properties are derived from open issues and neglected aspects identified from industrial IDS trends discussed in the previous section and IAS operational requirements. These industrial IDS trends have been summarized as follows:

- Predominance of Anomaly-based NIDS.
- Network Sensors that capture network traffic are strategically distributed across the system.
- Network Sensors forward captured data to a dedicated system for analysis.
- The system that analyzes data obtained from the Network Sensors is often centralized and located at the control, supervisory or business level of the automation hierarchy (Knapp, 2014). It may be deployed in specialized hardware or in a preexisting device in the automation environment that meets its computational requirements.

These tendencies convey that new components are integrated into the industrial system. These components communicate with one-another and consist of both hardware and software. Their integration adds new challenges, as their introduction into an automation system may affect its operation. In order to ensure that this does not occur, the operational requirements of the automation system must be considered. These requirements are capabilities that an IAS must possess in order to ensure its correct functionality. They are defined in (Stouffer et al., 2015) as the following: Real-time capabilities (OR1), High Availability (OR2), High Reliability and Fault Tolerance (OR3), High Maintainability (OR4) and Constrained Resources in certain embedded devices (OR5).

Considering the aforementioned trends and the operational requirements of IAS, the following properties for industrial IDS have been derived from (McHugh, 2001; Oryspayuli, 2006; Milenkoski et al., 2015; Zarpelão et al., 2017):

- *High Detection Accuracy (P1)*: Precise identification of intrusions. This means that an intrusion is identified as an intrusion, which would require a response. Whereas a non intrusive event is not identified as an intrusion.
- *Completeness (P2)*: Detection of a wide range of intrusions.
- *Real-time Intrusion Detection and Response or Timeliness (P3)*: Immediate detection and response in the presence of an intrusion. In other words, the time elapsed between the occurrence of an intrusion and its detection and response is as close to zero as possible.
- *Low Resource Consumption (P4)*: The amount of computational resources (i.e., energy consumption, network, CPU, ROM and RAM usage) used by industrial IDS components are the necessary to carry out their operation based on their required functionality and features. This means that the resource consumption is optimized in order to liberate computational resources that could be used by other system components.
- *Low Performance Overhead (P5)*: Direct or secondary effects of an IDS do not negatively influence the correct operation of the IAS. Considering the current industrial IDS trends, this indicates the following. Network Sensors do not add additional overhead to the industrial network. This is especially important, as the network load in industrial networks is high. It also entails that software components of the industrial IDS, that are not installed on specialized IDS hardware but rather on other IAS devices do not negatively influence the operation carried out by other automation components.
- *High Processing Performance (P6)*: Computational operations carried out by IDS components are performed in the shortest amount of time possible while maintaining reliable and high quality

results. Considering the current tendencies of industrial IDS, this indicates that IDS components are capable of real-time and reliable processing of highs amounts of data.

- *Fault Tolerance (P7)*: The IDS is capable of providing a degree of protection or functionality even in the presence of faults in its components (Lussier et al., 2004).
- *Robustness (P8)*: The IDS is capable of maintaining acceptable operation even in the presence of unexpected or adversarial circumstances that disturb its operation(Lussier et al., 2004).
- *Resiliency (P9)*: The IDS is capable of recovering from unexpected or adversarial circumstances that may have caused disturbances in its operation (Zhu and Basar, 2015).
- *Data Validity (P10)*: Data processed by the IDS is error-free. It is also authenticated and protected against unauthorized modifications by third-parties which could possibly invalidate the results of the industrial IDS.

An IDS possessing these properties ensures the fulfillment of IAS operational requirements as follows. High Detection Accuracy (*P1*), Completeness (*P2*), Real-Time Intrusion Detection (*P3*), High Processing Performance (*P6*) and Data Validity (*P10*) guarantee that intrusions that may negatively affect the correct operation of the automation system are correctly identified in order to implement timely countermeasures that may decrease or stop their effects. Hence, ensuring High Availability (*OR2*) and maintaining High Reliability (*OR3*). Furthermore, Low Resource Consumption (*P4*) and Low Performance Overhead (*P5*) ensure industrial IDS do not negatively influence the resources required to achieve RT Capabilities (*OR1*). They also open the opportunity for components of industrial IDS to be deployed in resource constrained devices (*OR5*). Moreover, High Maintainability (*OR4*) of the automation system is possible thanks to the correct and timely identification of intrusions (*P1* and *P3*) that guarantee the execution of actions to decrease, stop or repair the consequences of such intrusions. It is also supported by the Fault Tolerance (*P7*), Robustness (*P8*) and Resiliency (*P9*) of industrial IDS, which ensure the continuous operation of intrusion detection functions even in the presence of disturbances caused by faults or adversarial behavior. This is especially important in the presence of targeted attacks against the industrial IDS. Consequently, they provide continuous protection to the IAS that ensure its Availability and Reliability (*OR2* and *OR3*). An overview of this is provided in Table 1.

Table 1: Industrial IDS properties and their relation to operational requirements of Industrial Automation Systems (IAS).

IDS Properties	OR				
	1	2	3	4	5
P1-High Detection Accuracy	–	X	X	X	–
P2-Completeness	–	X	X	–	–
P3-RT ID and Response	–	X	X	X	–
P4-Low Resource Consumption	X	–	–	–	X
P5-Low Performance Overhead	X	X	–	–	X
P6-High Processing Performance	–	X	X	–	–
P7-Fault Tolerance	–	X	X	X	–
P8-Robustness	–	X	X	X	–
P9-Resiliency	–	X	X	X	–
P10-Data Validity	–	X	X	–	–

OR1: Real-time capabilities; **OR2**: High Availability; **OR3**: High Reliability and Fault Tolerance; **OR4**: High Maintainability; **OR5**: Constrained Resources.

Other properties such as interoperability, scalability and extendability are not discussed in detail in this contribution. This occurs, as these properties can not be evaluated or validated with metrics or specific tests, but rather depend on the specific features that are integrated into industrial IDS. On the other hand, the aforementioned properties are general attributes or characteristics that should be considered for industrial IDS with characteristics as those identified during the analysis of trends. Additional properties may be considered in order to validate more specific approaches or techniques. However, this is out of the scope of this contribution.

4 METRICS AND TESTS TO EVALUATE AND VALIDATE IDS PROPERTIES

This section presents the taxonomy of metrics and tests to evaluate and validate the aforementioned IDS properties. These properties comprise the dimensions of this taxonomy. Both metrics and tests allow to quantitatively and qualitatively measure them.

The metrics allow to quantitatively evaluate certain properties. They have been classified according to the data source of the values used for their calculation. These classes are: detection-, time-, computational resources- and IDS capacity metrics. This classification allows to identify which metrics can be analyzed based on the available system or test information. They have been abstracted from (McHugh, 2001; Kwon et al., 2015; Al-Jarrah et al., 2018; Buczak and Guven, 2016; Almalawi et al., 2013; Gupta and Chow, 2010; Brahmi et al., 2015; Leversage and Byres, 2008; Mitchell and Chen, 2014).

On the other hand, tests describe events or situations under certain conditions and environments (Athanasiades et al., 2003) where the behavior of components is analyzed or measured, which allows to use them as benchmarks from which metrics or conclusions can be obtained. These tests have been abstracted from the analysis of literature reviewed in Section 2 and key contributions (Puketza et al., 1996; Milenkoski et al., 2015; Durst et al., 1999).

4.1 Detection Metrics

Detection metrics are calculated from values resulting from the identification or misidentification of intrusions performed by an IDS and are often obtained from IDS tests, as it is necessary to know the total amount of intrusive and non-intrusive events that are fed to the IDS. They are comprised of the following:

- **True Positive (TP):** Number of intrusion events identified as intrusions.
- **False Negative (FN):** Number of intrusion events not identified as intrusions.
- **True Negative (TN):** Number of non-intrusive events not identified as intrusions.
- **False Positive (FP):** Number of non-intrusive events identified as intrusions.

From these values, a set of evaluation metrics can be obtained. Something characteristic that these metrics have is that they are used in machine learning to evaluate algorithms used for classification problems. This occurs as intrusion detection is inherently a classification problem.

Standard Detection Metrics: The following metrics are the most commonly used to evaluate IDS.

- **True Positive Rate (TPR):** It is also known as Sensitivity, Recall or Detection Rate (DR). Ratio of real intrusions detected from the total amount of existing intrusions. See (1).
- **True Negative Rate (TNR):** It is also known as Specificity. Ratio of real non-intrusive events identified as such from the total amount of non-intrusive events. See (1).
- **False Positive Rate (FPR):** It is also known as False Alarm Rate (FAR) or Fall-out. Ratio of non-intrusive events incorrectly identified as intrusions, from the total amount of non-intrusive events. See (2)
- **False Negative Rate (FNR):** Ratio of undetected intrusions from the total amount of real intrusions. See (2).

$$TPR = \frac{TP}{TP+FN} \quad TNR = \frac{TN}{TN+FP} \tag{1}$$

$$FPR = \frac{FP}{FP+TN} \quad FNR = \frac{FN}{FN+TP} \tag{2}$$

Precision (Pr): Ratio of correctly identified intrusions from all detected intrusions (i.e., regardless of whether or not they are really intrusive or non-intrusive events). It is also known as Positive Predictive Value (PPV) (Milenkoski et al., 2015).

$$Pr = \frac{TP}{TP+FP} \tag{3}$$

Accuracy (ACC): Ratio of correctly classified intrusive events and non-intrusive events considering all the classified events.

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

F-score: Harmonic mean of precision and recall (i.e., DR or TPR) (Almalawi et al., 2013). It represents the fraction of the detected intrusions that are valid.

$$F-score = 2\cdot\frac{Pr\cdot DR}{Pr+DR} = \frac{2\cdot TP}{2\cdot TP+FP+FN} \tag{5}$$

G-mean: Geometric mean of precision and recall (i.e., DR or TPR). A high G-mean indicates a high accuracy to detect intrusions.

$$G-mean = \sqrt{Pr\cdot DR} \tag{6}$$

Matthews Correlation Coefficient (MCC): Correlation coefficient between the observed and detected intrusions used when the difference between the amount of samples of two classes is too big (e.g., too many non-intrusive events and extremely few intrusions). Its value is within the range of -1 (total misidentification of intrusion) and 1 (perfect detection). A value of 0 represents a random identification (Al-Jarrah et al., 2018).

$$MCC = \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{7}$$

ROC Graph: The Receiver Operating Characteristic (ROC) graph allows to analyze the efficiency of the detection rate. The values commonly plotted are TPR vs FPR. Numerically, the area under the ROC curve is considered (Buczak and Guven, 2016).

4.2 Timing Metrics

Timing metrics are calculated from values resulting from the measurement of the temporal duration of events. These events include events related to the IDS, intrusions or other components that may interact with- or that exist within the same environment as the IDS.

Training Time (TT): Time required for an anomaly-based IDS to learn the expected normal behavior.

Intrusion Duration: Time elapsed between the start and end of an intrusion.

Detection Time (DT): Time elapsed between the start of an intrusion until it is detected by the IDS.

Response Time (RsT): Time elapsed between the detection of an intrusion until its response action is carried out.

Delay Time: Difference between the expected time of an event and the later time in which it really occurs. An example of this is the network delay. Network delay refers to the packet delivery delay in an industrial network (e.g end-to-end delay (Gupta and Chow, 2010)). This delay may be caused by IDS, an intrusion or other network components.

Processing Time per Event: Time elapsed since an event has started to be processed by an IDS, until a final decision regarding that event is made (i.e., whether or not it is an intrusive event).

4.3 Computational Resources Metrics

The following metrics quantify computational resources of certain system components. These components may be part of an IDS or they may exist in the same environment as one.

Percentage of Resource Utilization: Amount of a computational resource (per unit or percentage) used over a period of time. The resources that can be measured are **Network**, **CPU** and Memory (i.e., **RAM** & **ROM**) **utilization**.

Network Bandwidth: Amount of data (per unit or percentage) transmitted over the network (Brahmi et al., 2015).

4.4 IDS Capacity Metrics

The following metrics measure specific abilities of an IDS that are not related to detection accuracy.

IDS Throughput: Amount of events that can be processed by an IDS during a given time. In NIDS, it refers to the network traffic that can be processed by its Sensors and centralized analysis component.

Mean Time-to-Compromise (MTTC): Security metric that estimates the time required by an attacker to successfully impact a system (Leversage and Byres, 2008). Different methodologies exist to estimate this metric (McQueen et al., 2006; Leversage and Byres, 2008; Nzoukou et al., 2013).

4.5 Intrusion Identification Tests

Tests that verify the accuracy of an IDS (i.e., ability to detect intrusions (Puketza et al., 1996)). In order to provide validity to these tests, it is recommended that data fed into an IDS has similar characteristics to data that will be analyzed by it during normal operation. The use of predefined datasets to evaluate the accuracy of multiple intrusion detection approaches is a common practice (Milenkoski et al., 2015; Mitchell and Chen, 2014; Puketza et al., 1996), as it provides reproducibility. However, what type of- and how data is fed into the IDS depends on the type of IDS and the main approaches implemented in it.

In anomaly-based approaches, datasets are comprised of normal and abnormal behavior. Unfortunately, methodologies used to generate data with abnormal behavior may influence the quality of datasets and hence, the validity of the test itself. This has occurred with the dataset of the 1999 International Knowledge Discovery and Data Mining Competition (KDD99) and the Defense Advanced Research Projects Agency (DARPA) dataset DARPA98. Both datasets have been considered as unsuitable (McHugh, 2000). An alternative to these datasets are the Industrial Control System Cyber Attack Datasets (ICSCAD). These datasets have been popular for the study of cyber attacks in IAS. Their description is presented in (Morris and Gao, 2014). They consist of four datasets extracted from a gas pipeline and water storage tank testbeds.

On the other hand, signature-based approaches are more straightforward, as only the intrusions must be fed to the IDS.

4.6 Resource Usage Tests

Tests that verify the amount of computational resources used by IDS components (i.e., Network Sensors and centralized analysis component), as it is important to evaluate how much CPU and memory utilization is required. In (Milenkoski et al., 2015), two different approaches are identified. The first approach refers to the calculation of the overall resources required by an IDS. The second approach refers to the calculation of the resources required by individual components of an IDS. In both instances, it is necessary to perform the tests in both optimal and nonoptional conditions (i.e., together with stress tests), as the resources used by the IDS may variate depending on these conditions. It is also crucial to perform these tests with different configuration settings. This is especially important when the industrial IDS components are expected to be deployed on embedded hardware, which may require additional considerations (e.g., task configuration in Real-Time Operating Systems). The results of these tests allow to define the hardware requirements for IDS components.

4.7 Resource Overhead Tests

Tests that verify the degradation that may be caused by IDS. This degradation is observed during the decrease of performance of other system components that exist within the same environment. This performance may consist of metrics such as computational resources metrics, timing metrics or other specific metrics relevant to the system component being tested for degradation. These performance metrics are obtained under two different conditions (i.e., with and without an operational IDS). Performing these tests with different IDS configuration settings allows to identify appropriate configuration parameters for it.

4.8 Intrusion Detection Throughput Tests

Tests that identify the amount of workload that can be handled by IDS components (i.e., Network Sensors and centralized analysis component). The workload in Network Sensors is the amount of network traffic that is captured, preprocessed and forwarded to the centralized analysis component. On the other hand, the workload of the centralized analysis component is the amount of data received from the Network Sensors that can be processed. It is desirable that these tests are carried out under optimal and non-optimal conditions (similar to the resource usage tests).

4.9 Stress Tests

Tests that verify the IDS degradation under stressful situations created by external influences such as third-parties targeting IDS components (i.e., Network Sensors and centralized analysis component). This degradation can be observed during the IDS operation with the decrease in intrusion detection metrics (e.g., accuracy) and IDS capacity metrics; as well as the increase in time and computational resource utilization metrics. The most clear example of stressful situations are targeted attacks (Durst et al., 1999). In the field of IAS system faults may also generate stress.

5 TAXONOMY

The metrics and tests presented in the previous section evaluate and validate the IDS properties discussed in Section 3. However, not all metrics and tests are suitable for all properties. For this reason, a correlation between metrics and tests; and the IDS properties that they evaluate and validate is presented. An overview of this correlation and the suggested taxonomy is presented in Table 2. The correlation between tests, metrics and properties is measured in three different degrees: high, medium and no correlation.

A high correlation (i.e., ++) provides a clear and self-standing portrayal of the property. This means that a metric or test can provide by itself a solid evaluation or validation of a property. On the other hand, a medium correlation (i.e., +) provides semi-clear portrayal of the property. This means that for a metric or test to evaluate or validate a property, it is necessary to analyze their results with additional information (e.g., context, other metrics or tests, etc.). Furthermore, no correlation (i.e., –) indicates that a metric or test is not relevant for the evaluation or validation of a property.

Intrusion Detection Metrics grant an straightforward estimation of the High Detection Accuracy (*P1)* of an IDS. They are obtained from a *Intrusion Identification Test* and also provide a notion of Completeness (*P2*). Which is validated when a variety of different intrusions are used for testing. *Stress Tests* can also be used as they may contain complex adversarial models.

Time Metrics obtained during *Intrusion Identification Test* provide an estimation of Timeliness (*P3*). Detection Time (*DT*), Response Time (*RsT*) and *Processing Time per Event* provide a clear assessment of the RT capabilities of an IDS. Training Time (*TT*) provides a notion on usability on the time capabilities of an IDS (e.g., how fast can the IDS start operating).

Furthermore, *Network, Memory and CPU Utilization Metrics* provide an straightforward estimation of the Low Resource Consumption (*P4*) and Low Performance Overhead (*P5*) of an IDS. These metrics are obtained from *Resource Usage Tests*. In addition to these metrics and test, some *Time and IDS metrics* are considered to evaluate the Performance Overhead (*P5*). *Resource Overhead, Intrusion Detection Throughput and Stress Tests* may be used to obtain these metrics. In addition, increase in *Delay Time* observed in other components, as well as the decrease in IDS *Throughput* also provide an estimate of overhead in the system. Together with the previously mentioned metrics, Detection Time (*DT*) and Response Time (*RsT*) can also be analyzed to find appropriate configuration settings for the IDS, which would allow to decrease its performance overhead.

Intrusion Detection and Time Metrics can be used as a measure of quality for High Processing Performance (*P6*), Fault Tolerance (*P7*), Robustness (*P8*) and Resiliency(*P8*), as they provide a view of the IDS

Table 2: Taxonomy of metrics and tests to evaluate and validate industrial IDS properties.

Metric & Tests			P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Metrics	Detection	Standard Metrics	++	+	—	—	—	+	+	+	+	—
		Precision (Pr)	++	+	—	—	—	+	+	+	+	—
		Accuracy (Acc)	++	+	—	—	—	+	+	+	+	—
		F-score	++	+	—	—	—	+	+	+	+	—
		G-mean	++	+	—	—	—	+	+	+	+	—
		MCC	++	+	—	—	—	+	+	+	+	—
		ROC Graph	++	+	—	—	—	+	+	+	+	—
	Time	Training Time (TT)	—	—	+	—	—	++	—	—	—	—
		Intrusion Duration	—	—	—	—	—	—	+	+	+	—
		Detection Time (DT)	—	—	++	—	+	++	+	+	+	—
		Response Time (RsT)	—	—	++	—	+	++	+	+	+	—
		Delay Time	—	—	—	—	++	—	+	+	+	—
		Processing Time per Event	—	—	++	—	+	++	+	+	+	—
	C. Rsrc.	Network Utilization	—	—	—	++	++	—	—	—	—	—
		Memory Utilization	—	—	—	++	++	—	—	—	—	—
		CPU Utilization	—	—	—	++	++	—	—	—	—	—
		Network Bandwidth	—	—	—	++	++	—	—	—	—	—
	Cap.	Throughput	—	—	+	—	+	++	+	+	—	—
		MTTC	—	—	—	—	—	—	++	++	+	++
Tests		Intrusion Identification	++	+	+	—	—	+	+	+	+	—
		Resource Usage	—	—	—	++	++	—	—	—	—	—
		Resource Overhead	—	—	—	—	++	—	—	—	—	—
		ID Throughput	—	—	—	—	++	—	++	++	++	—
		Stress	—	+	—	—	+	—	++	++	++	++

IDS Properties: **P1**: High Detection Accuracy; **P2**: Completeness; **P3**: Real-time Intrusion Detection; **P4**: Low Resource Consumption; **P5**: Low Performance Overhead; **P6**: High Processing Performance; **P7**: Fault Tolerance; **P8**: Robustness; **P9**: Resiliency; **P10**: Data Validity.
Correlation: **++**: High; **+**: Medium; **–**: No Correlation. Other: **C**: Computational; **Rsrc.**: Resources; **Cap.**: Capacity.

performance. However, these metrics by themselves are not enough to evaluate these properties.

A more accurate estimation of High Processing Performance (*P6*) can be obtained by considering the Training Time (*TT*), Detection Time (*DT*), Response Time (*RsT*), *Processing Time per Event* and IDS *Throughput*. As an IDS with a high processing performance has the most accurate detection in the shortest amount of time possible.

Besides *Intrusion Detection Metrics, Timing metrics and Intrusion Detection Throughput* and *Stress Tests* support the assessment of Fault Tolerance (*P7*), Robustness (*P8*) and Resiliency (*P9*). A fault tolerant, robust and resilient IDS maintains a high accuracy and real-time capabilities in the presence of faults or other unexpected circumstances. It also has a behavior as close to the optimal one it was designed for (i.e., IDS *Throughput*). Moreover, it has a high *MTTC*, which means that an adversary should require more time in order to compromise the system. The evaluation of the *MTTC* allows to identify mechanisms to harden or protect the IDS. This provides a notion of Data Validity (*P10*).

6 CONCLUSIONS

The protection of IAS is critical to ensure their correct operation. Industrial IDS are a feasible solution that can be seamlessly integrated in them for this, as they can detect a wide range of intrusions. However, before this integration occurs it is necessary to verify that an IDS is efficient, reliable and does not negatively influence the operation of the target system (i.e., does not violate the operational requirements). In order to do this, a qualitative and quantitative evaluation of its performance must be made. The outcome of this evaluation is to verify the suitability of IDS for its deployment in an automation environment. Unfortunately, from the analysis of literature performed in this contribution; it was observed that the quality of experimental evaluations from consulted IDS has often been overlooked (Tavallaee et al., 2010).

To address this issue, a taxonomy of metrics and tests that allow to evaluate and validate properties of an industrial IDS has been proposed. First, a set of open issues in IDS in both the computer science and automation field have been identified. The identification of these open issues has been supported by the consideration of operational requirements for IAS, which allowed to define properties that an IDS should have in order to be suitable for an automation environment. Afterwards, a set of metrics and tests to evaluate and validate these properties have been abstracted from reviewed literature. Furthermore, the presented taxonomy allows to identify appropriate metrics and tests to be considered during the development of an industrial IDS depending on its desired properties.

The presented taxonomy focuses on analyzing the overall performance of an industrial IDS and its influence over the automation system. Aspects related to specific implementation approaches or techniques are not considered, as these may require additional tests and metrics. Some examples of this are: performance metrics for Machine Learning algorithms (Servin and Kudenko, 2008), entropy metrics for anomaly detection (Marnerides et al., 2014), protocol-specific metrics (Kwon et al., 2015), etc.

REFERENCES

Al-Jarrah, O. Y., Al-Hammdi, Y., Yoo, P. D., Muhaidat, S., and Al-Qutayri, M. (2018). Semi-supervised multi-layered clustering model for intrusion detection. *Digital Communications and Networks*, 4(4):277–286.

Almalawi, A., Tari, Z., Fahad, A., and Khalil, I. (2013). A framework for improving the accuracy of unsupervised intrusion detection for scada systems. In *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 292–301.

Athanasiades, N., Abler, R., Levine, J., Owen, H., and Riley, G. (2003). Intrusion detection testing and benchmarking methodologies. In *Proceedings First IEEE International Workshop on Information Assurance (IWIAS 2003)*, pages 63–72. IEEE.

Bhuyan, M. H., Bhattacharyya, D. K., and Kalita, J. K. (2014). Network anomaly detection: Methods, systems and tools. *IEEE Communications Surveys & Tutorials*, 16(1):303–336.

Brahmi, I., Brahmi, H., and Yahia, S. B. (2015). A multi-agents intrusion detection system using ontology and clustering techniques. In Amine, A., Bellatreche, L., Elberrichi, Z., Neuhold, E. J., and Wrembel, R., editors, *Computer Science and Its Applications*, volume 456 of *IFIP Advances in Information and Communication Technology*, pages 381–393. Springer International Publishing, Cham.

Buczak, A. L. and Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2):1153–1176.

Durst, R., Champion, T., Witten, B., Miller, E., and Spagnuolo, L. (1999). Testing and evaluating computer intrusion detection systems. *Communications of the ACM*, 42(7):53–61.

Garitano, I., Uribeetxeberria, R., and Zurutuza, U. (2011). A review of scada anomaly detection systems. In Kacprzyk, J., Corchado, E., Snášel, V., Sedano, J., Hassanien, A. E., Calvo, J. L., and Ślzak, D., editors, *Soft Computing Models in Industrial and Environmental Applications, 6th International Conference SOCO 2011*, volume 87 of *Advances in Intelligent and Soft Computing*, pages 357–366. Springer Berlin Heidelberg, Berlin, Heidelberg.

Gupta, R. A. and Chow, M.-Y. (2010). Networked control system: Overview and research trends. *IEEE Transactions on Industrial Electronics*, 57(7):2527–2535.

Kleinmann, A. and Wool, A. (2017). Automatic construction of statechart-based anomaly detection models for multi-threaded industrial control systems. *ACM Transactions on Intelligent Systems and Technology*, 8(4):1–21.

Knapp, E. D. (2014). *Industrial network security: Securing critical infrastructure networks for smart grid, scada, and other industrial control systems*. Elsevier, Waltham MA, 2nd edition edition.

Kwon, Y., Kim, H. K., Lim, Y. H., and Lim, J. I. (2015). A behavior-based intrusion detection technique for smart grid infrastructure. In *2015 IEEE Eindhoven PowerTech*, pages 1–6. IEEE.

Leversage, D. J. and Byres, E. J. (2008). Estimating a system's mean time-to-compromise. *IEEE Security & Privacy Magazine*, 6(1):52–60.

Littler, T., Wang, H. F., Yang, Y., McLaughlin, K., and Sezer, S. (2013). Rule-based intrusion detection system for scada networks. In *2nd IET Renewable Power Generation Conference (RPG 2013)*, pages 1–4. IET.

Lussier, B., Chatila, R., Ingrand, F., Killijian, M. O., and Powell, D. (2004). On fault tolerance and robustness in autonomous systems. In *the Third IARP-IEEE/RAS-EURON Joint Workshop on Technical Challenges for Dependable Robots in Human Environments 2004*, pages 1–7. Citeseer.

Maglaras, L. A. and Jiang, J. (2014). Ocsvm model combined with k-means recursive clustering for intrusion detection in scada systems. In *2014 10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine)*, pages 133–134.

Mahan, R. E., Fluckiger, J. D., Clements, S. L., Tews, C. W., Burnette, J. R., Goranson, C. A., and Kirkham, H. (2011). Secure data transfer guidance for industrial control and scada systems: Pnnl-20776.

Marnerides, A. K., Schaeffer-Filho, A., and Mauthe, A. (2014). Traffic anomaly diagnosis in internet backbone networks: A survey. *Computer Networks*, 73:224–243.

McHugh, J. (2000). Testing intrusion detection systems: A critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory. *ACM Transactions on Information and System Security*, 3(4):262–294.

McHugh, J. (2001). Intrusion and intrusion detection. *International Journal of Information Security*, 1(1):14–35.

McQueen, M. A., Boyer, W. F., Flynn, M. A., and Beitel, G. A. (2006). Time-to-compromise model for cyber risk reduction estimation. In Gollmann, D., Massacci, F., and Yautsiukhin, A., editors, *Quality of Protection*, volume 23 of *Advances in Information Security*, pages 49–64. Springer US, Boston, MA.

Milenkoski, A., Vieira, M., Kounev, S., Avritzer, A., and Payne, B. D. (2015). Evaluating computer intrusion detection systems: A survey of common practices. *ACM Computing Surveys*, 48(1):1–41.

Mitchell, R. and Chen, I.-R. (2014). A survey of intrusion detection techniques for cyber-physical systems. *ACM Computing Surveys*, 46(4):1–29.

Morris, T. and Gao, W. (2014). Industrial control system traffic data sets for intrusion detection research. In Butts, J. and Shenoi, S., editors, *Critical Infrastructure Protection VIII*, volume 441 of *IFIP Advances in Information and Communication Technology*, pages 65–78. Springer Berlin Heidelberg, Berlin, Heidelberg.

Naedele, M. and Biderbost, O. (2004). Human-assisted intrusion detection for process control systems. In *Proceedings of the Second International Conference on Applied Cryptography and Network Security*, pages 216–225. Citeseer.

Nivethan, J. and Papa, M. (2016). A scada intrusion detection framework that incorporates process semantics. In Trien, J. P., Prowell, S. J., Goodall, J. R., and Bridges, R. A., editors, *Proceedings of the 11th Annual Cyber and Information Security Research Conference*, pages 1–5. ACM.

Nzoukou, W., Wang, L., Jajodia, S., and Singhal, A. (2013). A unified framework for measuring a network's mean time-to-compromise. In *2013 IEEE 32nd International Symposium on Reliable Distributed Systems (SRDS)*, pages 215–224. IEEE.

Oryspayuli, O. D. (August 2006). *What intrusion detection approaches work well if only TCP/IP packet header information is available?* PhD thesis, Master Thesis, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands, Enschede, Netherlands.

Paxson, V. (1999). Bro: A system for detecting network intruders in real-time. *Computer Networks*, 31(23-24):2435–2463.

Paxson, V. (2007). Considerations and pitfalls for conducting intrusion detection research: Keynote. In *Fourth GI International Conference on Detection of Intrusions & Malware, and Vulnerability Assessment (DIMVA)*.

Perkins, E., Contu, R., and Alaybeyi, S. B. (2017). Market guide for operational technology security.

Ponomarev, S. and Atkison, T. (2016). Industrial control system network intrusion detection by telemetry analysis. *IEEE Transactions on Dependable and Secure Computing*, 13(2):252–260.

Puketza, N. J., Zhang, K., Chung, M., Mukherjee, B., and Olsson, R. A. (1996). A methodology for testing intrusion detection systems. *IEEE Transactions on Software Engineering*, 22(10):719–729.

Roesch, M. (1999). Snort - lightweight intrusion detection for networks. In *13th USENIX Conference on System Administration*, LISA '99, pages 229–238, Berkeley, CA, USA. USENIX Association.

Servin, A. and Kudenko, D. (2008). Multi-agent reinforcement learning for intrusion detection. In Tuyls, K., Nowe, A., Guessoum, Z., and Kudenko, D., editors, *Adaptive Agents and Multi-Agent Systems III. Adaptation and Multi-Agent Learning*, volume 4865 of *Lecture Notes in Computer Science*, pages 211–223. Springer Berlin Heidelberg, Berlin, Heidelberg.

Sommer, R. and Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy*, pages 305–316. IEEE.

Stouffer, K., Pillitteri, V., Lightman, S., Abrams, M., and Hahn, A. (2015). *Guide to Industrial Control Systems (ICS) Security: Supervisory, Control and Data Acquisition (SCADA) Systems, Distributed Control Systems (DCS) and Other Control System Configurations such as Programmable Logic Controllers (PLC): NIST Special Publication 800-82*. National Institute of Standards and Technology, revision 2 edition.

Tavallaee, M., Stakhanova, N., and Ghorbani, A. A. (2010). Toward credible evaluation of anomaly-based intrusion-detection methods. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(5):516–524.

Udd, R., Asplund, M., Nadjm-Tehrani, S., Kazemtabrizi, M., and Ekstedt, M. (2016). Exploiting bro for intrusion detection in a scada system. In Zhou, J. and Lopez, J., editors, *the 2nd ACM International Workshop on Cyber-Physical System Security*, pages 44–51. ACM.

Valdes, A. and Cheung, S. (2009). Communication pattern anomaly detection in process control systems. In *2009 IEEE Conference on Technologies for Homeland Security (HST)*, pages 22–29. IEEE.

Yang, Y., McLaughlin, K., Sezer, S., Yuan, Y. B., and Huang, W. (2014). Stateful intrusion detection for iec 60870-5-104 scada security. In *2014 IEEE PES General Meeting*, pages 1–5. IEEE.

Zarpelão, B. B., Miani, R. S., Kawakani, C. T., and de Alvarenga, S. C. (2017). A survey of intrusion detection in internet of things. *Journal of Network and Computer Applications*, 84:25–37.

Zhang, J., Gan, S., Liu, X., and Zhu, P. (2016). Intrusion detection in scada systems by traffic periodicity and telemetry analysis. In *2016 IEEE Symposium on Computers and Communication (ISCC)*, pages 318–325. IEEE.

Zhu, B. and Sastry, S. (2010). Scada-specific intrusion detection/prevention systems: a survey and taxonomy. In *Proceedings of the 1st Workshop on Secure Control Systems (SCS)*, volume 11, pages 1–7.

Zhu, Q. and Basar, T. (2015). Game-theoretic methods for robustness, security, and resilience of cyberphysical control systems: Games-in-games principle for optimal cross-layer resilient control systems. *IEEE Control Systems*, 35(1):46–65.