Importance of Machine Learning Techniques to Improve the Open Source Intrusion Detection Systems

Fatimetou Abdou Vadhil, Mohamedade Farouk Nanne, Mohamed Lemine Salihi
Department of Mathematics and Computer Science, University of Nouakchott Al-assriya, Mauritania

Article Info

Article history:

Received Jul 14, 2021
Revised Aug 19, 2021
Accepted Sep 17, 2021

Keywords:

IDS
SIEM
Open Source
Machine Learning
ELK

ABSTRACT

Nowadays, it became difficult to ensure data security because of the rapid development of information technology according to the Vs of Big Data. To secure a network against malicious activities and to ensure data protection, an intrusion detection system played a very important role. The main objective was to obtain a high-performance solution capable of detecting different types of attacks around the system. The main aim of this paper is to study the lacks of traditional and open source Intrusion Detection Systems and the Machine Learning techniques commonly used to overcome these lacks. A comparison of some existing works by Intrusion Detection System type, detection method, algorithm and accuracy was provided.

Corresponding Author:

Fatimetou Abdou Vadhil,
Department of Mathematics and Computer Science,
University of Nouakchott Al-assriya,
Nouakchott, Mauritania.
Email: fatiab38@gmail.com

1. INTRODUCTION

The amount of data processed and stored on personal, industrial and government digital networks is constantly increasing, which strongly motivates attackers to make illegitimate access attempts. The protection of personal and sensitive data and the security of the digital environment should be addressed progressively in order to respond to the challenges linked to modern security concerns with regard to log management and the ability to define the extent of security policy violations.

The open source and commercial tools that can be used against security threats are enormous and their use should be context and need-specific. This study focuses on Intrusion Detection Systems (IDS) and Security Information and Event Management (SIEM). SIEM manages the data generated by the IDSs. We have chosen the most popular traditional and open source IDSs that have scored high marks in intrusion detection such as Snort, Suricata, Bro (Zeek) and OSSEC (Open Source SECurity). The first motivation to use these open source IDSs is their ability to detect known attacks effectively with little consumption of computational resources and little time consuming. However, the number of false alarms generated by these IDSs is actually a challenge.

Another challenge related to the traditional IDSs is that they cannot detect previously unknown attacks, because they are based on signatures or rules, and detect only based on the defined rules. Additionally, sometimes these IDSs have difficulty detecting variations in known attacks according to Y. Ding et al. [1]. In addition, most of these IDSs do not have Graphical User Interface (GUI).

In order to address the false alarms problem, the simplest way is to turn off some attack signatures; however, this can degrade the detection quality of the IDS according to B. Subba et al. [2].

N. Hubballia et al. [3] present a survey of false alarm minimization techniques in signature-based IDS such as Signature Enhancement, Stateful Signatures, Vulnerability Signatures, Alarm Correlation, Alarm Verification etc. The advantages and drawbacks of each of these techniques were provided. Ref. [3] give also an analysis on commercial SIEMs that uses some of the techniques presented in this survey. B. Subba et al. [2] mentioned many drawbacks related to false alarm minimization existing techniques.

Another technique that often gives good results in terms of detection rate and can achieve false alarm reduction in a proactive manner is Machine Learning. In addition, Meng et al. [4] prove that using machine learning to create a false alarm filter is a promising solution to solve the false alarm problem.

Some research papers studied the application of these techniques to improve IDSs by making them more accurate in recognizing malicious network traffics. Machine Learning-based IDSs was the object of a survey conducted by Kunal et al. [5]; they provide a comparison of some previous works based on type of classifier, approach and dataset used and the results obtained in each work.

Gilmore and Haydaman [6] presented a taxonomy of available Machine Learning methods, and highlighted the advantages and weaknesses of each. Furthermore in order to improve intrusion detection, a number of interesting observations which provide insight into the application of Machine Learning techniques to IDSs has been revealed by [7] during an experimental analysis.

Among the most common ideas to meet the challenges of open source IDSs is to create a filter that serves to minimize false alarms, as proposed in papers [8] and [9].

2. SECURITY INFORMATION AND EVENTS MANAGEMENT

An SIEM (according to Gartner [10]) is a solution for threat detection and security incident response through real-time collection and historical analysis of security events from a wide variety of events and contextual data sources; it also supports compliance reporting and incident investigation by analyzing historical data from these sources.

New types of attacks and vulnerabilities are discovered daily. Firewalls, IDS, IPS and other security solutions designed for malicious activity at various locations in the IT infrastructure. However, many solutions on the market are not effective and are not even capable of detecting unknown attacks, which may reflect that these solutions do not consider context. It is essential for a security system to understand the context of the activities to be secured. Analyze and monitor the traffic that passes through the system's infrastructure in order to know all the information surrounding it. This is where SIEM is useful.

An excellent example is the Gartner reports, which present a detailed assessment of current SIEM systems based on multiple characteristics depending on the context [11]. Taking into account the characteristics and indicators is therefore an important task that must precede the selection. Authors in [12] suggest appropriate technological and operational requirements that they have found useful in an SIEM system and proposes a two-phase evaluation process to measure the compliance and applicability of an SIEM.

The determining characteristics of an SIEM are described below. The features described are basic features covered by the majority of commercial and open source SIEM systems: Data collection, Normalization and categorization, Notifications, Correlation, Visualization, Prioritization, Reporting and Workflows.

Podzins and Romanovs [13] present the advantages and disadvantages of SIEM. Among the main problems of deployment of the SIEM solution presented in this article are the following:

- The need for a high level of maintenance to investigate alerts and optimize SIEM (correction of "false positives") will quickly become overwhelming if care is not taken.
- SIEM will not provide complete information without other security solutions such as firewalls, IPS / IDS and other security solutions.

2.1. Commercial SIEMs

Commercial SIEMs offer full features. In addition to a SIEM having pre-integrated intrusion detection, vulnerability analysis and behavior-based monitoring system capabilities, the majority of these systems measure events based on the number of events received per second (EPS).

Large companies only trust commercial systems, never thinking that some systems development researchers consider flaws found in commercial systems in the development of open source systems. We can get along with the study made by [14] which provide a comparison study of some commercial SIEMs such as Splunk (this one is the leader since 2012, according to Gartner Magic Quadrant), QRadar, LogRhythm and ArcSight Enterprise Security Manager (ESM) according to the following criteria:

Real-time monitoring, Threat intelligence, Behavior profiling, Data and user monitoring, Application monitoring, Analytics, Log management and reporting, Deployment/Support Simplicity.

Another aspect that many companies are targeting when using commercial systems is compliance with privacy laws such as GDPR (General Data Protection Regulation), CNIL (Commission Nationale Informatique et Liberté), etc. However, most high-performance SIEM systems are very expensive, so few companies have the capacity to use them. Moreover, the implementation of an open source system with similar features is a challenge [18].

In addition, Son and Kwon [15] finds that commercial SIEMs often have limited flexibility for optimization. Flexibility allows easy adaptation to the conditions and environment. Maintenance also plays an important role in this context. Developing one's own SIEM can be a flexible and less expensive solution.

2.2. Open Source SIEMs

Open source SIEMs are many, and of different levels. In this paper, we cite those that are considered the most effective in terms of performance in intrusion detection. These are Security Onion, Graylog, but also ELK (Elasticsearch, Logstash and Kibana) [16], according to Gantner [17]. There is also OSSIM which is among the most popular open source SIEMs and uses many open source tools, small and large, to achieve the desired feature. However, this is not a cure for all kinds of threats [18].

Authors in Ref. [19] analyze and compares tested SIEMs solutions (OSSIM, ELK Stack, Splunk Free and Graylog), their features and their compliance with GDPR. In this work, four SIEMs were analyzed and, based on the results obtained from the tests, systematized in the two tables presented, it was concluded that the OSSIM and Splunk Free solutions are not scalable, so that the choice for the implementation of the prototype is between ELK Stack and Graylog. The two chosen SIEMs take into consideration the legal requirements of the GDPR, such as the anonymisation and pseudo-anonymisation of sensitive data, the retention time of the "logs", their encryption and their protection.

ELK Stack therefore has a feature that allows the pseudo-anonymisation of data, which is one of the fundamental requirements for acting in accordance with the GDPR. In addition, Graylog makes it possible to restrict user access to certain information.

2.3. ELK as a SIEM

ELK is a Big Data platform for log analysis and management, but also an alternative to commercial SIEMs. ELK is chosen among the 12 SIEM solutions presented by Gartner. It is an extremely popular log management and analysis tool and has the advantage of being free and open source. A large community is approaching Open Source products. In Ref. [20], the results of the SIEM evaluation tests performed by these authors show that performance and flexibility are not necessarily linked to commercial SIEM. Even so, there are difficulties related to commercial IDSs. Juniper, Palo Alto and Check Point are the leading commercial IDS/IPS on the market, a survey was conducted to evaluate their detection capabilities using three datasets, the results obtained show that the detected attacks do not exceed 50% of all attacks [21].

Therefore, it is better to use an open source SIEM, which will allow more freedom on the modification or integration of new features if necessary. These integrations are often necessary to fill the gaps in feature often missing in open source solutions.

To use ELK as a security tool, [22] proposed to combine intrusion detection systems with Machine Learning techniques for intrusion detection and network security alerting using Elasticsearch as the core for data storage. For this purpose, some ML-jobs have been integrated into the ELK.

On other hand, [23] provide a survey on intrusion detection using Deep Learning technique and how a Deep Learning model could be integrated into the ELK Stack.

3. INTRUSION DETECTION SYSTEMS

Intrusion Detection Systems (IDS) are security tools that aim to defend a system, execute countermeasures or generate alerts to an entity capable of performing appropriate actions, when an attack occurs [24]. The notions SIEM and IDS are strongly linked. SIEMs have been designed for the first time to reduce the number of False Positives generated by Intrusion Detection Systems (IDS).

These systems could also be focused on a variety of areas. Some IDSs act as advanced firewalls and detect attacks on network entrances, others could monitor the network internally to catch intruders, or even collect information about the entire network for central analysis. Most of these systems have a similar structure and set of features [25], as seen in Figure 1.

3.1. Data Source

There are two types of detection:

- Host-based Intrusion Detection (HIDS): Monitors the characteristics of a single host and the events occurring in that host to detect suspicious activity.
- Network based Intrusion Detection (NIDS): Monitors network traffic for particular network segments or devices and analyzes the network.

These two types can be combined to find a Hybrid- based Intrusion Detection, which is the combination of HIDS and NIDS. HIDSs and NIDSs are generally complementary in a malicious activity detection system.

3.2. Detection Methods

Intrusion detection systems can be classified under four methods:

- Signature-based Intrusion Detection: It is a methodology for detecting hosts and malicious network activity based on known malicious patterns or sequences.
- Anomaly-based Intrusion Detection: Anomaly-based detection shows abnormal or anomalous system behavior. It creates the profile of normal activities, if the normal activity exceeds the given threshold, it is considered as an intrusion. Any deviation from the threshold, gives the abnormal behavior.
- Hybrid-based Intrusion Detection: These systems typically use signature-based detection for normal traffic. It is a combination of the two approaches above (Signature-based and Anomaly-based) to avoid the disadvantages and to integrate the advantages.
- Protocol-based Intrusion Detection: It is a method for monitoring the protocols used by the system while performing state and dynamic behavior analysis and applying the legal use of the protocol.

3.3. Structure

- Centralized (or monomod): Single-mode IDSs are deployed separately in stand-alone mode and not all devices/applications communicate with each other.
- Distributed: multiple instances (sensors) of cooperative IDS that are configured and controlled by a centralized IDS server.

3.4. Response Type

- Active: Active IDS is known as an intrusion prevention system (IPS) [26].
- Passive: It can generate or log alerts in a file only after an anomaly has been identified [27].

3.5. Frequency usage

- Real Time: Real-time systems detect abnormal behavior as it occurs.
- Offline: these systems process recorded attack data sets.

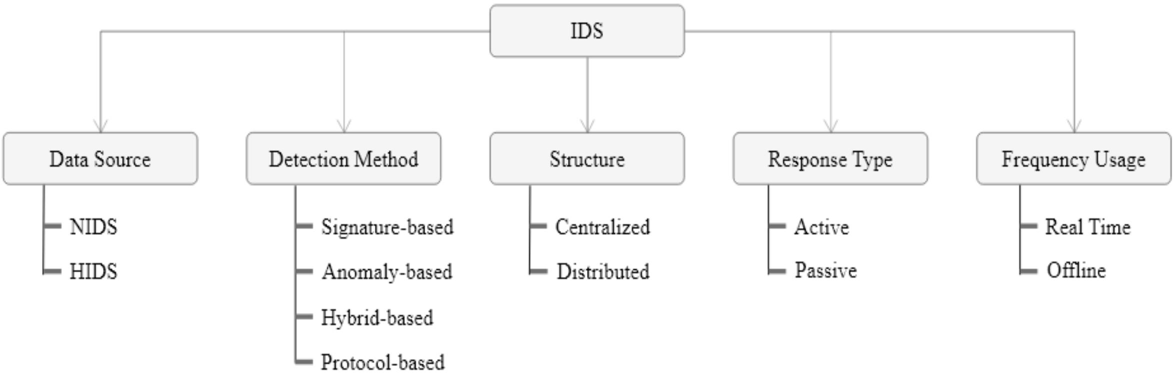

Figure 1. Intrusion Detection Systems Taxonomy

4. OPEN SOURCE IDS

For this study, we selected the most commonly used IDSs such as Snort, Zeek and Suricata as NIDS (Network-based IDS) and OSSEC as HIDS (Host-based IDS).

4.1. Snort

Snort [28] is an open source intrusion detection and prevention tool. It is a single process, which means that a single job can be run in a session without interruption. Snort uses only intrusion detection based on user signatures and the community maintains the rules. It tries to match each packet it receives with a set of rules defined by the Snort configuration (rules can be specified to detect certain contents in the packet payload or other characteristics found in packet headers). Snort is rule-based, when a rule is compared to a packet, Snort can take actions such as Alert, Log packet, Ignore packet and Drop packet. Snort has managed to be the most popular IDS in various aspects thanks to its scalability. However, it still generates a high number of False-Positive alerts. It is unable to detect unknown [29]. A solution named SNIPER [30] using the "few-shot learning" method to minimize the FP rate can be associated with Snort to solve this problem.

4.2. Suricata

Just like Snort, Suricata [31] is a free and open source intrusion detection and prevention tool, it supports additional features such as Multi-Process Analysis, which provides distributed analysis of large volumes of data. The list of additional features is quite long. Suricata also supports automatic protocol detection for specific protocols on all ports. This feature limits the amount of configuration required for the solution to perform basic features. The contributions to this NIDS are limited, there is [32] who has designed a framework to use Suricata rules as official rules rather than implementing independent preprocessors or detection engines. By this design, it is possible to implement state analysis methods by defining state rules without major modification of the plugin.

4.3. Zeek (Formally "Bro")

Zeek [33] is an open source NIDS that uses behavioral analysis to detect a network anomaly. It allows network administrators to perform incident response, forensic analysis, file extraction and hashing. Zeek is an advanced tool that captures metadata about network activity and then provides an interpreter to understand the activity [34]. It supports a wide range of inbound traffic analysis, even outside the security domain, including troubleshooting and performance measurement. There are many significant advantages to using Zeek, for example, it efficiently captures data from Gbps networks and can operate very efficiently in a high-speed environment. It is well known for its flexibility to customize feature. However, it has some limitations such as the difficulty of deployment [35].

4.4. OSSEC

OSSEC [36] is an open source HIDS whose response type is Active that uses both hybrid anomaly detection methodologies. It is capable of operating system log analysis, integrity checking, Windows registry monitoring, active response and real-time alerts. It enables multi-system monitoring due to its centralized and multi-platform architecture. The IDS Log Analysis Engine is capable of correlating and analyzing logs from multiple hosts. The implementation of this IDS is being studied in open source platform development communities and has been considered in Wazuh [37], which can be considered as an enhanced extension.

Smart SIEM [20] combines ELK with Snort, Bro and OSSEC, it is a good way to take advantage of a 100% open source SIEM based on a Big Data platform and intrusion detection systems, which are ranked among the most powerful open source IDSs. On the other hand, the integration of these IDSs as is, without addressing their failures (false alarms, memory consumption, etc.) is to be discussed.

5. MACHINE LEARNING-based IDS

In SIEM and IDS, in general, two types of Machine Learning algorithms are used: Supervised classification for abuse detection and Non-supervised outlier/novelty classification for anomaly detection [38]. Machine Learning algorithms are starting to be used more and more to improve and make IDSs flexible. As quoted by [39], the commonly algorithms integrated with detection systems:

Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearst Neighbor (KNN), Naïve Bayes, Logistic regression, Decision tree, K-means, Deep Brief Network (DBN), Deep Neural Network (DNN), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Generative Adversarial Network (GAN), Restricted Boltzmann Machine (RBM), Auto Encoder.

Authors in [40] propose a NIDS based on Fuzzy-Genetic and Genetic Algorithm to efficiently detect various types of network intrusions. On the other hand, problems and challenges related to Machine Learning for the detection and prevention of network attacks are discussed in [41]. Therefore, it was discovered that Machine Learning techniques play an important role in security area, and that these techniques have the potential to cause a low false alarm rate while providing a high detection rate, these are what the majority of Open Source systems need.

Several works use Machine Learning and Data Mining techniques to maximize the efficiency of intrusion identification after generating data using Snort. Attack detection is complementary to Snort, as long as Snort only detects known attacks. Suthaharan [42] proposed advanced rules for Snort, designed to detect new attacks and reduce false positives rate, to achieve this purpose. The test of this method was achieved using a Data Mining package named "Weka", the used algorithms are: KNN, Random Forest, ID3, Adaboost, Multi-Layer Perceptron MLP, Naïve Bayes, Quadratic Discriminant Analysis QDA and j48 with CICIDS dataset.

In the context of high speed networks, Ref. [43] show that, Snort priorities of true positive traffic can be approximated in real-time by a decision tree classifier, using the information of easily extracted features. He executed Snort on all the ISCX dataset, extracted all packets generating and triggering Snort alerts, then, he mapped between these packets and their flows in the ISCX dataset and label different alerts based on the ISCX labeled flows. By the aforementioned steps, he constructed a labeled IDS dataset that contain 7779 alerts. To make these tasks, he used several tools include TCPReplay and TCPDump. Features extracted were protocol, source IP, source port, destination IP and destination port. Afterwards the Decision Tree model is applied to Dataset. He obtained an accuracy of 99%, while avoiding its false alerts, and being able to run in real-time.

A hybrid NIDS proposed by authors in Ref. [44] to maximize the effectiveness in identifying attacks by integrating the Network Traffic Anomaly Detection (NETAD) [45] as an anomaly-based IDS with SNORT. Afterwards, k-means and CART algorithms to classify normal and abnormal traffics. Then, the proposed hybrid IDS is evaluated using KDD Cup Dataset.

Another contribution to Snort is a new plug-in [46] developed to address false alarm problems. Three datasets (NSA, DARPA and NSL-KDD) were selected to conduct performance experiments on Machine Learning algorithms. The evaluation environment for the experiment was built and consisted of a configuration and data pre-processing using Weka. Then, a set of high-performance Machine Learning algorithms was selected based on the results of previous research and guidelines: Support Vector Machine (SVM), Decision Trees (DT), Fuzzy Logic, BayesNet and NaiveBayes. The best result was obtained using SVM optimized with a False Positive Rate equal to 8.6% and a False Negative Rate equal to 2.2%.

Some authors believe that reducing the amount of incoming data could help reduce false alarms, as the idea evoked by [47]. The idea of Ref. [47] is to develop a model to reduce the amount of data to be processed by IDS, using a flow-based approach using Bro IDS based on some algorithms available on Weka library such as; J48, Random Tree, Rep Tree (Reduced-error pruning Tree), BF Tree (Best-First tree), PART, Jrib (JRepeated Incremental Pruning), DTNB (Decision Table Naïve Bayes) for classification. This approach has generated a significant number of false positive alarms. This indicates (according to the author of this paper) that for detection purposes, it is difficult to make a complete behavior of malicious activities from limited data and flow level.

On the other hand, Gustavsson [48] uses six supervised Machine Learning algorithms (Support Vector Machines SVM, Naive Bayes NB, Quadratic Discriminant Analysis QDA, Artificial Neural Networks ANN, Decision Tree DT and Random Forest RF) on Zeek logs to improve malicious traffic detection, templates and scripts have been created to extract the necessary feature using a dataset labeled.

Another hybrid NIDS was proposed by [49] to detect attacks in the network by monitoring network traffic. This NIDS aims to detect and stop attacks in real time impairing the security of the LAN, it uses Suricata as signature-based detection to discover known attacks, and the Isolation Forest algorithm (an unsupervised Machine Learning algorithm) to detect a network anomaly. The authors of this paper believe that by applying Suricata before the Isolation Forest algorithm, the latter should only detect unknown attacks. Another contribution to Suricata is presented in [50], which describes a project called OPNids integrates Suricata with the "DragonFly" Machine Learning engine, which uses a continuous data analysis model to ingest Suricata's line-rate network data and help make decisions. There is a commercial version of OPNids under development.

The work on improving OSSEC using Machine Learning is not much. As a contribution to OSSEC, there is [51], which provides a User Behavior Analysis based OSSEC using a Naive Bayes algorithm. The paper use OSSEC as a HIDS for monitoring of user shell commands and detects intrusion based on those commands. The detection system is based on Naive Bayes model. This solution detect intrusions with an accuracy of 69%. They used UNIX-User to train their model.

Additionally, a recent IDS [52] has scored significant points of effectiveness in terms of attack detection and false alarm reduction, which has been compared with Snort, the proposed IDS is called "INsIDES" and uses Machine Learning to achieve more effective attack detection than Snort. The proposed IDS is compared to Snort using the new UNSW-NB15 dataset. The results of this comparison show a detection rate of 98.11% and a false alarm rate of 8.57% for INsIDES, while Snort has a detection rate of 2.43% and a false alarm rate of 30.66%, showing that traditional IDSs can integrate Machine Learning techniques well. However, this IDS does not have the ability to detect unknown attacks.

Table 1 shows some a summary of works that contributes to open source IDSs using Machine Learning techniques with some results. For propositions that provide more than one result, we have selected the highest accuracy and precision for each.

Other IDS solutions were provided in different context, some papers use the context of IoT such as the papers [8], [53] and [54]. Ref. [54] examines the possibilities of using Machine Learning algorithms to protect the IoT against DoS attacks. Classifiers are the subject of an in-depth study that can advance the development of anomaly-based intrusion detection systems. Authors in [54] place particular emphasis on the evaluation of the performance of supervised ML algorithms, they find that the use of unsupervised ML can be more efficient, so they tend to evaluate the performance of unsupervised ML algorithms for the detection of intrusions in the IoT will be taken into account in their future work.

In this regard, Naukarkar and Hande [55] proposed an IDS model using supervised Machine Learning approach. The proposed approach identify the attack by analyzing information from the KDD Cup dataset. It use Naive Gaussian Bayes algorithm to classify the traffic data generated. The NSL-KDD and UNSWNB15 dataset were used to assess the efficiency and effectiveness of an intrusion detection approach based on Machine Learning algorithms using a Data Mining tool [56].

Table 1. Contributions to traditional IDSs using Machine Learning

Paper	Journal, Conference., Thesis, Project	Detection Type	Detection Method	Proposition	ML Techniques	Datasets	Accuracy	Precision
[42]	Journal	NIDS	Anomaly-based	Developed advanced rules for Snort	KNN, RF, ID3, Adaboost, MLP, NaiveBayes, QDA, J48	CICIDS	N/A	98%
[43]	Journal	NIDS	Anomaly-based	Real-time attack detection on high-speed traffic	Decision Tree	ISCX	99.3%	98.45%
[44]	Journal	NIDS	Hybrid-based	Effective Intrusion Identification and False Alarm Elimination	K-means, Decision Tree	KDD Cup 99	99.41%	99.40%
[46]	Journal	NIDS	Anomaly-based	Snort False Alarm Reduction Plug-in	SVM, DT, Fuzzy Logic, BayesNet, NaiveBayes	NSA, DARPA, NSL-KDD	95.6%	N/A
[47]	Journal	NIDS	Anomaly-based	Develop a model to reduce the amount of data to be processed through intrusion detection	J48, Random Tree, Rep Tree, BF Tree, PART, Jrib, DTNB	ISOT, CTU-50, CTU-51, CTU-52, CTU-53	N/A	66%
[48]	Thesis	NIDS	Anomaly-based	Create templates and scripts to improve traffic detection	SVM, NB, QDA, ANN, DT et RF	CICIDS2017	99.16%	N/A
[49]	Conference	NIDS	Hybrid-based	Design of hybrid IDS combining Suricata with Isolation Forest	Isolation Forest	N/A	N/A	N/A
[50]	Project	NIDS	Signature-based	Help improve incident response and threat-hunting activities	Dragonfly	N/A	N/A	N/A
[51]	Project	HIDS	Hybrid-based	Implement User Behavior Analysis on cloud infrastructures	Naive Bayes	UNIX-User	69.1%	78%

🟦 Works on Snort 🟨 Works on Zeek (Bro) 🟫 Works on Suricata 🟩 Works on OSSEC

Most of the previous research is limited to the supervised learning. The unsupervised learning is little used compared to the supervised learning; the cause of this comes down to some challenges. This type of learning was used by [57] through the Auto-Encoder algorithm. The proposed framework was tested using the CICIDS2017.

However, whatever the type of learning used for the detection, a very important step must be part of the creation of IDS is the Feature Selection. Feature selection aims to select the most relevant features and eliminate the other features, which reduces dimensionality and model learning time and improves detection results. Features selection methods are classified into three major categories, Filter, Wrapper, and Embedded [58].

Kurniabudi et al. [59] use Information Gain (which is a Filter-based feature selection method) to select relevant features, they implements five models to evaluate its performance, namely Random Forest, Bayes Net, Random Tree, Naive Bayes and J48. To test this approach, authors use CICIDS-2017 dataset. They obtained good results in terms of accuracy and execution time. The best models were Random Forest with an accuracy of 99.86% and 22 features and J48 with an accuracy of 99.87% and 52 features.

D. Stiawan et al [60] introduced an approach for constructing ensemble IDS using six ranked feature selection techniques, namely, Information Gain, Gain Ratio, Symmetrical Uncertainty, Relief-F, One-R and Chi-Square ensemble with four classifiers such as Bayesian Network, Naïve Bayesian, J48 and SOM, and validated using Hold-up, K-fold approaches. Experimental results were obtained on Weka using the ITD-UTM dataset. The highest accuracy was 85.2593% obtained using Bayesian Network and Symmetrical Uncertainty with 10 features.

6. CONCLUSION

The aim of this paper is to present a study on open source IDSs, emphasizing their strengths and weaknesses. Snort, Suricata, Zeek and OSSEC IDSs were discussed.

Some works that provided contributions to improve these open source IDSs were also presented. Most of the proposed solutions use supervised Machine Learning algorithms, and find that they tends to make the open source IDSs more efficient in terms of the accuracy of detection and the minimization of false alarms. Existing works that propose new ideas to improve Snort, Suricata, Zeek and OSSEC and the results they have achieved by integrating Machine Learning techniques in these IDS are presented.

Regarding SIEM, a solution like ELK Stack, free, open source, scalable and recognized for its performance in data management and analysis as well as its ease of use, deserves to be chosen to manage the data generated by the IDSs.

REFERENCES

[1] Y. Ding, M. Xiao and A. Liu, "Research and implementation on snort-based hybrid intrusion detection system," *2009 International Conference on Machine Learning and Cybernetics*, vol. 3, pp. 1414-1418, 2009.

[2] B. Subba, S. Biswas and S. Karmakar, "False alarm reduction in signature-based IDS: game theory approach," *Security and Communication Networks*, vol. 9, pp. 4863-4881, 2016.

[3] N. Hubballia and V. Suryanarayananb, "False Alarm Minimization Techniques in Signature-Based Intrusion Detection Systems: A Survey," *Computer Communications*, vol. 49, pp. 1-17, 2014.

[4] Y. Meng and L. Kwok, "Adaptive False Alarm Filter Using Machine Learning in Intrusion Detection." *Practical Applications of Intelligent Systems*, vol. 124 573–584, 2011.

[5] Kunal and M. Dua, "Machine Learning Approach to IDS: A Comprehensive Review," *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, pp. 117-121, 2019.

[6] C. Gilmore and J. Haydaman, "Anomaly Detection and Machine Learning Methods for Network Intrusion Detection: an Industrially Focused Literature Review," *Int'l Conf. Security and Management | SAM'16*, pp. 292-298, 2016.

[7] L. Zomlot, S. Chandran, D. Caragea and X. Ou, "Aiding Intrusion Analysis Using Machine Learning," *2013 12th International Conference on Machine Learning and Applications*, vol. 2, pp. 40-47, 2013.

[8] W. Meng, W. Li and L. F. Kwok, "EFM: Enhancing the performance of signature-based network intrusion detection systems using enhanced filter mechanism," *Computers & Security*, Vol. 43, pp. 189-204, 2014.

[9] Y. Meng and L. F. Kwok, "Adaptive False Alarm Filter Using Machine Learning in Intrusion Detection," *Practical Applications of Intelligent Systems*, vol.124, pp. 573–584, 2011.

[10] A. T. Williams and M. Nicolett, "Improve it security with vulnerability management," 2005.

[11] G. Sadowski, K. Kavanagh and T. Bussa, "Critical Capabilities for Security Information and Event Management," [Retrieved February 2020], Available from: https://www.gartner.com/en/documents/3981260/critical-capabilities-for-security-information-and-event .

[12] H. Mokalled, R. Catelli, V. Casola, D. Debertol, E. Meda and R. Zunino, "The applicability of a SIEM solution: Requirements and Evaluation," *IEEE 28th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pp. 132-137, 2019.

[13] O. Podzins and A. Romanovs, "Why SIEM is Irreplaceable in a Secure IT Environment?," *2019 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, pp. 1-5 , 2019.

[14] K. Agrawal and H. Makwana, "A Study on Critical Capabilities for Security Information and Event Management," *International Journal of Science and Research (IJSR)*, vol. 4, pp. 1893-1896, 2015.

[15] S. J. Son and Y. Kwon, "Performance of ELK stack and commercial system in security log analysis," *2017 IEEE 13th Malaysia International Conference on Communications (MICC)*, pp. 187-190, 2017.

[16] Elastic Stack, [Retrieved April 2020], Available from: https://www.elastic.co/

[17] Gartner, "Security Information and Event Management (SIEM Tools) Reviews," [Retrieved July 2019], Available from: https://goo.gl/U4UxGM

[18] D. Hermanowski, "Open Source Security Information Management system supporting IT security audit," *2015 IEEE 2nd International Conference on Cybernetics (CYBCONF)*, pp. 336-341, 2015.

[19] A. Vazão, L. Santos, M. B. Piedade and C. Rabadão, "SIEM Open Source Solutions: A Comparative Study," *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pp. 1-5, 2019.

[20] M. Elarass, N. Souissi, "Smart SIEM: From Big Data logs and events to Smart Data alerts," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 8, pp. 3186-3191, 2019.

[21] D. Hedemalm, "An Empirical Comparison of the Market-Leading IDS's," *Thesis at the University of Halmstad*, 2018.

[22] O. Negoita and M. Carabas, "Enhanced Security Using Elasticsearch and Machine Learning," *Intelligent Computing*, vol. 1230, pp. 244–254, 2020.

[23] M. Raut, S. Dhavale, A. Singh and A. Mehra, "Insider Threat Detection using Deep Learning: A Review," *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 856-863, 2020

[24] S. Axelsson, "Intrusion detection systems: A survey and taxonomy," 2000.

[25] R. Robbins, "Distributed Intrusion Detection Systems: An Introduction and Review," *SANS Institute.* [Retrieved December 2019], Available from: https://www.sans.org/readingroom/whitepapers/detection/distributed-intrusion-detectionsystems-introduction-review-897

[26] R. F. Pratama, N. A. Suwastika and M. A. Nugroho, "Design and Implementation Adaptive Intrusion Prevention System (IPS) for Attack Prevention in Software-Defined Network (SDN) Architecture," *2018 6th International Conference on Information and Communication Technology (ICoICT)*, pp. 299-304, 2018.

[27] R. Kumar and D. Sharma, "HyINT: Signature-Anomaly Intrusion Detection System," *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pp. 1-7, 2018.

[28] Snort website, Available from: https://www.snort.org

[29] M. Roesch, "Snort- Lightweight Intrusion Detection for Networks," *13th Systems Administration Conference- LISA '99: Proceedings of the 13th USENIX conference on System administration*, pp. 229–238, 1999.

[30] Y. Koizumi, S. Murata, N. Harada, S. Saito and H. Uematsu, "SNIPER: Few-shot Learning for Anomaly Detection to Minimize False-negative Rate with Ensured True-positive Rate," *ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 915-919, 2019.

[31] Suricata website, Available from: https://suricata-ids.org

[32] B. Kang, K. McLaughlin and S. Sezer, "Towards A Stateful Analysis Framework for Smart Grid Network Intrusion Detection," *4th International Symposium for ICS & SCADA Cyber Security Research 2016 (ICS-CSR)*, pp. 124-131, 2016.

[33] Zeek website, Available from: https://docs.zeek.org/en/stable/intro/

[34] Z. Kai, "Research and Design of the Distributed Intrusion Detection System Based on Snort," *2012 International Conference on Computer Science and Electronics Engineering*, vol. 2, pp. 525-527, 2012.

[35] T. U. Sheikh, H. Rahman, H. S. Al-Qahtani, T. K. Hazra and N. U. Sheikh, "Countermeasure of Attack Vectors using Signature-Based IDS in IoT Environments," *IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 1130-1136, 2019.

[36] OSSEC website, Available from: https://www.ossec.net

[37] Wazuh website, Available from: https://wazuh.com/

[38] A. Sapegin, "High-Speed Security Log Analytics Using Hybrid Outlier Detection," *Thesis at the University of Potsdam*, 2019.

[39] H. Liu and B. Lang, "Machine Learning and Deep Learning Methods for Intrusion Detection Systems: A Survey," *Applied Sciences*, vol. 9, pp. 1-28, 2019.

[40] P. Mahadik and S. Gosavi, K. Varade and P. Sagare, "New Approach for Intrusion Detection System using Fuzzy Genetic Algorithm," 2015.

[41] S. Suthaharan, "Big Data Classification: Problems and Challenges in Network Intrusion Prediction with Machine Learning," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, pp. 70–73, 2014.

[42] N. I. Andrey and E. B. Michael, "Realization of Expert Intrusion Detection System Based on the Results of Datasets and Machine Learning Algorithm Analysis," *CASPIAN JOURNAL, Management and High Technologies*, vol. 2, pp. 100-107, 2020.

[43] A. Ammar, "A Decision Tree Classifier for Intrusion Detection Priority Tagging," *Journal of Computer and Communications*, vol. 3, pp. 52-58, 2015.

[44] J. Patel and K. Panchal, "Effective Intrusion Detection System using Data Mining Technique," *Journal of Emerging Technologies and Innovative Research*, Vol. 2, pp. 1869-1878, 2015.

[45] M. V. Mahoney, "Network Traffic Anomaly Detection Based on Packet Bytes," 2003, [Retrieved July 2019], Available from : https://cs.fit.edu/~mmahoney/paper6.pdf

[46] S. A. Shah and B. Issac, "Performance Comparison of Intrusion Detection Systems and Application of Machine Learning to Snort System," *Future Generation Computer Systems*, vol. 80, pp. 157-170, 2018.

[47] H. Alaidaros and M. Mahmuddin, "Flow-Based Approach on Bro Intrusion Detection," *Journal of Telecommunication, Electronic and Computer Engineering*, Vol. 9, pp. 2-2, 2017.

[48] V. Gustavsson, "Machine Learning for a Network-based Intrusion Detection System : An application using Zeek and the CICIDS2017 dataset," 2019.

[49] Z. Chiba, N. Abghour, K. Moussaid, A. El Omri and M. Rida, "Newest collaborative and hybrid network intrusion detection framework based on suricata and isolation forest algorithm," *Proceedings of the 4th International Conference on Smart City Applications*, 2019.

[50] S. M. Kerner, "OPNids Integrates Machine Learning Into Open-Source Suricata IDS," [Retrieved April 2019], Available from: https://www.eweek.com/security/opnids-integrates-machine-learning-into-open-source-suricata-ids

[51] R. Ravi and K. Dzeparoska, "VM User Behavior Analysis using OSSEC on SAVI Testbed," [Retrieved July 2019], Available from: http://rajsimmanravi.github.io/assets/files/Final_Project_Report.pdf

[52] L. A. Valero, "INsIDES: A new Machine Learning-based Intrusion Detection System," Available from: https://repositori.upf.edu/bitstream/handle/10230/32875/Valero_2017.pdf?isAllowed=y&sequence=1

[53] L. Santos, R. Gonçalves, and C. Rabadão, "A Novel Intrusion Detection System Architecture for Internet of Things Network," *European Conference on Cyber Warfare and Security*, pp. 428-435, 2019.

[54] A. Verma, and V. Ranga, "Machine Learning Based Intrusion Detection Systems for IoT Applications," *Wireless Personal Communications*, pp. 2287-2310, 2020.

[55] R. L. Naukarkar and K. N. Hande, "Analysis of Implementing Network Intrusion Detection (NIDS) Algorithms Using Machine Learning", *International Journal of All Research Writings*, Vol. 1, pp. 2582-1008, 2020.

[56] P. Maniriho, L. Mahoro, E. Niyigaba, Z. Bizimana and T. Ahmad, "Detecting Intrusions in Computer Network Traffic with Machine Learning Approaches", *International Journal of Intelligent Engineering and Systems*, Vol. 13, pp. 433-445, 2020.

[57] C. Zhang, Y. Chen, Y. Meng, F. Ruan , R. Chen, Y. Li and Y. Yang, "A Novel Framework Design of Network Intrusion Detection Based on Machine Learning Techniques," *Security and Communication Networks*, vol. 2021, pp. 1-15, 2021.

[58] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003.

[59] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Idris, A. M. Bamhdi and R. Budiarto, "CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection," *IEEE Access*, vol. 8, pp. 132911-132921, 2020.

[60] D. Stiawan, A. Heryanto, A. Bardadi, D. P. Rini, I. M. I. Subroto, Kurniabudi, M. Y. B. Idris, A. H. Abdullah, B. Kerim and R. Budiarto, "An Approach for Optimizing Ensemble Intrusion Detection Systems," *IEEE Access*, vol. 9, pp. 6930-6947, 2021.

BIOGRAPHY OF AUTHORS

PhD student at the University of Nouakchott Al-assriya, Research Unit: Scientific Computing, Computer Science and Data Science.
Research topic: The use of artificial intelligence for the improvement of open source security systems.

Teacher Researcher in the Mathematics and Computer Science department at the University of Nouakchott Al-assriya.
Axis of research: Artificial Intelligence, Network Security.

Teacher Researcher in the Mathematics and Computer Science department at the University of Nouakchott Al-assriya.
Axis of research: Scientific Computing, Artificial Intelligence.