Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 201 (2022) 205–212

Procedia Computer Science

www.elsevier.com/locate/procedia

The 13th International Conference on Ambient Systems, Networks and Technologies (ANT)
March 22-25, 2022, Porto, Portugal

Intrusion Detection Systems using Supervised Machine Learning Techniques: A survey

Emad E. Abdallah*, Wafa' Eleisah, Ahmed Fawzi Otoom

Faculty of Prince Al-Hussein Bin Abdullah II for Information Technology
The Hashemite University, P.O. Box 330127, Zarqa 13133, Jordan

Abstract

In this paper, we investigate the subject of intrusion detection using supervised machine learning methods. The main goal is to provide a taxonomy for linked intrusion detection systems and supervised machine learning algorithms. For this purpose, we provide a deep discussion of the concepts of intrusion detection systems, supervised machine learning techniques, and cyber-security attacks. Then, concerning the application of supervised learning for intrusion detection, we cover relevant efforts. Finally, a taxonomy is provided based on these related works. Based on this taxonomy, we can conclude that the classification performance of supervised learning algorithms is high and promising based on a study of four popular data sets in this domain: KDD'99, NSL-KDD, CICIDS2017, and UNSW-NB15. Moreover, feature selection is important and, in many cases, is needed for an enhancement in performance. Furthermore, data imbalance can be a concern, and sampling approaches can help resolve the issue. Finally, for good performance, large intrusion detection data sets necessitate a deep learning technique.

Keywords: intrusion detection systems, cyber security, supervised machine learning algorithms.

* Corresponding author. Tel.: Tel.: +0096-795673231; fax: +00962(05) 3826625.
E-mail address: emad@hu.edu.jo

10.1016/j.procs.2022.03.029

1. Introduction:

With the increase in internet usage, there is an increase in possible cyber-attacks. In many cases, these attacks are new and require intelligent systems for detecting them. An intrusion detection system (IDS) analyzes network traffic to identify any type of malicious traffic. Generally, IDS are divided into two main types: misuse-based IDS and anomaly-based IDS, which we will discuss deeply in the next section. In brief, misuse-based detection searches for existing attacks and matches the new traffic with these attacks, if there is a match, an alarm is raised. On the other hand, anomaly-based detection, search the new traffic for any deviation from the normal behavior and report it as an anomaly, or not normal. Anomaly-based detection is important for identifying zero-based attacks [24].

For a successful detection of new attacks, a huge amount of data must be used for building a model for what is normal and what is an anomaly. This raised attention to the application of supervised machine learning algorithms for the efficient understanding of the data and for building a predictive model that can predict new attacks with high-performance rates. The high dimensionality of the data is another issue, as the increase of the feature space with a relatively low number of data, (records) can lead to a "curse of dimensionality" problem that affects the results of classification, and this attracted attention for the application of feature selection and feature reduction for an enhancement of classification.

The high interest in the area of applying supervised machine learning for IDS motivated us to investigate more in this area for more understanding of the best approaches that can be implemented for the successful detection of attacks.

Hence, the main objective of this paper is to present a survey for supervised learning algorithms and intrusion detection systems. We review necessary concepts that are related to IDS, The concept of IDS is first presented, where we discuss its definition, types, and importance. Then, we review popular supervised learning algorithms and popular data sets in this domain and the concept of dimensionality reduction. Also, we provide a summary of cyber-security attacks. After the background section, we review deeply related works in the area of supervised machine learning and IDS. Finally, we present a taxonomy the can guide what is suitable for algorithms across different types of IDS datasets, and it guides whether feature selection is effective or not in improving classification performance.

This paper is organized as follows: Section 2 discusses the background part about IDS, supervised machine learning, and cyber-security attacks. Section 3 discusses the application of supervised machine learning algorithms for IDS. Section 4 presents the taxonomy. Section 4 summarizes the paper and draws up conclusions.

2. Background

This section provides a background on three important topics that are necessary for understanding the following sections of the paper: intrusions detection systems, supervised machine learning, and cyber security attacks.

2.1. Intrusion detection systems (IDS)

The connection of computers is important for communication and the exchange of information. However, computers can be exposed to external threats that need to be continuously monitored and detected [1]. These threats can violate the computer security triad: confidentiality, integrity, and availability (CIA).

An intrusion is an attempt to compromise the CIA or evade the security mechanism of a computer or a network. Intrusion detection is the process of monitoring and analyzing the traffic in a network or a computer for signs of intrusion [2]. The main objectives of the IDS can be summarized as 1) Monitoring hosts and networks, 2) Analyzing the behaviors of computer networks, 3) Generating alerts, 4) Responding to suspicious behaviors.

Many IDS systems exist in the literature. However, many IDS suffer from two problems:

1. High false alarm rate: where alarms are raised from non-threatening violations and many serious violations are left undetected.
2. New attacks are not easily detected which raised the attention toward the use of modern machine learning techniques for detecting intrusions.

Generally, IDS can be classified based on two ways: classification based on methodology, and 2) classification based on the monitored platform. In the following subsections, we will discuss these classifications in more detail.

IDS can be classified based on methodology into anomaly detection and misuse-based detection. Anomaly detection methods model the user behavior to build a normal (standard) behavior. This method observers the behavior of a user over a period of time and tries to build a model that is close to the normal behavior of the user. Whenever there is an event that deviates from the normal behavior, the IDS detects it as suspicious behavior and raises an alarm which makes it suitable for detecting zero-day attacks. Two major disadvantages are related to this model: 1) the inability to deal with the evolving of normal user behavior, and 2) a high false-positive rate [4].

On the other hand, in the misuse detection or signature-based method, captured events are compared with known attacks or threats for the detection of possible intrusions. The known attacks are often referred to as, signatures [5]. This method is more accurate and raises fewer false alarms compared to the anomaly detection method. However, it is not useful for detecting new attacks [4].

IDS can also be classified based on the monitored platform into two types: host-based IDS (HIDS) and network-based IDS (NIDS). HIDS monitors and analyzes the activities on the system (host) where it is deployed [6]. It is capable of monitoring parts of the dynamic behavior and state of the system. On the other hand, NIDS monitors network traffic to detect remote attacks, attacks that are carried over a network connection [6]. It is a device that is distributed within networks to inspect traffic crossing the devices it sits on. It can be a hardware or software-based system. Usually, NIDS have two network interfaces: one used to listen to network conversations, and the other one is used for control and reporting [7].

2.2. Supervised Machine learning techniques

There are two different machine learning techniques [8-11] that can be applied for the purpose of automatic detection of intrusions: supervised learning and unsupervised learning. The main focus of this survey is the application of supervised machine learning techniques for IDS. In supervised learning or classification, there must be labeled data the can be used for training a model for detection purposes. The process of classification can be summarized in the following steps:

- Data collection, where data is collected to provide the necessary information for training the classification model. Usually, the data set is characterized using a feature set that can be discriminative between the classes. The collection of data is not an easy task, and hence, several benchmark data sets exist such as KDD'99 [14] and NSL-KDD [14], and UNSW-Nb15 [15] and CICIDS2017 [14]. Table 1 illustrates a brief description of each data set.

Table 1. Description of IDS data sets

Dataset	Description
KDD'99	It is generated using simulation of normal and attacks traffic in a military environment (US AirForce LAN). It contains nine weeks of simulation in rat tcpdump files. The dataset is characterized using 41 features related to intrinsic, content, and traffic. Four types of attacks are simulated: DoS, Prob, U2R, and R2L.
NSL-KDD	It is a modification to the KDD'99 dataset with solving the problems of redundancy, duplicates, the imbalance of data.
UNSW-Nb15	It was created using the IXIA PefectStorm tool to extract normal and attack network traffic based on 100 GB of raw network traffic. It is characterized using 49 features. It consists of around 175 thousand records for training and around 82 thousand records for testing. There are nine types of attacks: Fuzzers, Analysis, Backdoor, DoS, Exploit, Generic, Reconnaissance, Shellcode, Worm
CICIDS2017	It was created in an emulated environment in a 5 day period. It contains traffic in packet flow and bidirectional flow. 80 features are extracted. Attacks involve: Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet, and DDoS

- Data reduction: the high dimensionality of the feature space can be problematic and can lead to problems such as the "curse of dimensionality", where there is a relatively low number of training data in a very high dimensional space [13]. That is why data can be transformed into a lower-dimensional space using methods like Principle component analysis (PCA) or linear discriminant analysis (LDA) [8]. In other direction, best performing features can also be selected using a feature selection algorithm such Best First algorithm [9].

- Classification: At this stage, Part of the data is used to build the model (training) and another part is used to test the performance of the classification model (testing). A higher number of machine learning algorithms exist in the literature for building the model and will discuss popular types of them, later in this section. The division of data into a training set and testing set only is referred to as a hold-out test. Another common type of test is the N-fold cross-validation test, where the data is divided into N folds, nine of them are used for training and the tenth fold is used for testing the model. Then, another set of nine folds is used for training and the tenth fold is used for testing, and the process repeats. The accuracy is calculated as the average accuracy across all folds.
- Performance evaluation: the performance of the classification model is evaluated and, usually, the evaluation is done using accuracy and false positive rate (FPR)

2.3. Cyber-security attacks:

IDS can be deployed for the detection of a variety of attacks. According to [16], cyber security attacks can be categorized based on purpose, legal classification, based on severity of involvement, based on scope, and based on network types.

The attacks that are based on purpose include reconnaissance attacks, access attacks, and denial of service attacks. A reconnaissance attack is a dangerous type of attack as the attacker trap victims into becoming their friend to extract sensitive information from them [17]. Examples of these attacks include packet sniffers, scanning the port, and queries regarding internet information. In access attacks, the intruder has the ability to gain access to a device. Examples of these types of attacks include man-in-the-middle attacks, phishing, social engineering, and attacks on secret code [16]. The third type of attack with this category is a denial of service (DoS) attack. In a DoS attack, the attacker exploits the connectivity of the internet to cripple the services offered by the victim site, simply, by flooding a victim site with a high number of requests. It can be a single source attack or a multi-source attack, where, in the latter, it is referred to as distributed denial of service attack (DDoS) [18]. Examples of DoS attacks include Smurf, SYN flood, and DNS attacks.

In the second type of categorization, legal classification attacks, the attacks include cybercrime, cyber espionage, cyber terrorism, and cyberwar. Cybercrime attacks example is identity theft which involves the use of an account without the owner's permission [19]. Another type within this category is cyber espionage, or cyberspying attack, which involves the use of computer networks for gaining illegal access to confidential information especially associated with governments [20]. Cyber terrorism attacks are carried by extremists using cyberspace. Finally, cyber wars are the use of cyberspace for conducting wars between nations.

In the third type of categorization, attacks can be classified based on the severity of involvement into two types: active attacks and passive attacks [16]. Simply, the difference between these two types of attacks is that, in the first attack, the attacker aims at altering system resources or modifying its operation, whereas, in the latter type, the attacker makes use of the information without any alteration or modification of resources or operations. Examples of active attacks include spoofing, man-in-the-middle attacks, buffer overflow, and others. Passive attacks examples are keystroke logging [21] and Backdoors.

Cyber-attacks can be classed into malicious and non-malicious attacks in the fourth category of categorization. Malicious assaults use various sorts of software, such as viruses, worms, Trojan horses, spyware, adware, botnets, and other types, to carry out an attack with the goal of causing harm. Non-malicious assaults, on the other hand, are unintentional attacks carried out by untrained staff that may result in modest data loss [22].

The final type of categorization of cyber security attacks is based on network types where attacks are classified according to the network types such as mobile ad hoc networks (MANET) and wireless sensor networks (WSN) [23]. Examples of attacks on MANET include black hole attack, flood rushing attack, Byzantine attack. Other examples on WSN include application-layer attacks, network layer attacks, and other network layers attacks [16].

3. Machine learning and IDS

Recently, there has been extensive research in the application of supervised machine learning techniques for automating the intrusion detection process.

For example, the authors in [12], targeted the problem of classifying intrusions using the NSL-KDD data set. First, the data set is preprocessed to solve missing data issues and to categorize numeric attributes. Then, the data set is clustered into four data sets and partitioned into training and test data. Then, the data is fed to a random forest classifier, and classification and accuracy, and FPR are calculated. A feature selection approach is also applied using the Symmetrical Uncertainty measure. The authors reported a slight enhancement in performance after the application of feature selection. The reported results are compared with that of C4.5, but random forests outperformed C4.5 [12]. On average, after feature selection, the reported accuracy is 99.67 and the false alarm rate is .005.

Similar to the author in [12], the authors in [24] examined the performance of ten classification algorithms on the NSL-KDD dataset, with a different feature selection approach to that of [12]. The feature selection approach is based on the application of attribute evaluators and filtering. The authors implemented the following algorithms: Naive Bayes, Bayes-Net, Logistics, Random tree, Random forest, J48, Bagging, OneR, PART, ZERO. The best performing classification algorithm is a random forest with an accuracy of 99.9% and a low false alarm rate of 0.001. The second-best performing classifier is Bagging with an accuracy of 99.8% and a similar performance is reported by the PART algorithm.

Similarly, the authors in [25] targeted the NSL-KDD data set and applied a hold-out testing approach with no use of a feature selection approach. The four classification algorithms are tested: random forest, SVM, logistics regression, and Gaussian mixture model. The random forest proved to be the best algorithm with a reported accuracy of 99%. The second-best performing classifier is logistics regression with a reported accuracy of 84%.

Recently, the author in [26] experimented with the performance of artificial neural networks (ANN) and SVM on a sample of the NSL-KDD data set. The sample represents 20% of the whole data set. Two feature selection methods are used: Correlation-based and Chi-square-based methods. The first resulted in a selection of 17 features and the latter resulted in 35 features. After feature selection, the data is fed to ANN and SVM classifiers. The results using correlation-based feature selection and ANN reported the highest performance with 94.0% accuracy.

In another direction, the KDD'99 data set was used by the authors in [27] where the classes are: normal, Prob, DoS, U2R, and R2L. The feature selection algorithms CFSSubSet Eval and Best First are used in combination with four methods: one unsupervised method, k-means, and three supervised methods, SVM, Naïve Bayes, and random forest. The three supervised methods outperformed the unsupervised technique and are based on the eight best features. The best-reported accuracy is the random forest with an accuracy of 99.0%.

The KDD'99 was also used by the authors in [28], where the Ant Colony Algorithm is first applied for choosing a suitable representative set of the original dataset with 550 samples (records). Then, the authors applied a novel method for feature reduction named the Gradual Feature Removal (GFR) method to reduce the dimensionality of the feature space into 19 features. After that, the reduced features are combined with SVM for classification. The reported accuracy is 98.67% before feature selection and there has been no real improvement of accuracy after feature selection and the results are 98.62%.

In [29], the authors examined the effect of reducing the dimensional space of the KDD'99 on the overall classification performance. They examined the feature selection algorithm, information gain based on entropy (IG), and combined the resulted feature vector (22 feature) with Back Propagation Neural Network (BBNN). The results showed no changes in accuracy after the feature reduction as it remained 91%.

Another data set that attracted attention within this domain is the CICIDS2017 dataset. For example, the authors in [30, 31], implemented an approach based on sampling, feature reduction, and boosting for building an IDS. The dataset has a major problem related to the imbalance of the dataset. Thus, they pre-processed the data using Synthetic Minority Oversampling Technique (SMOT) to deal with low numbers of instances in some classes. Then, they applied a reduction approach based on ensemble feature selection (EFS) and principal component analysis (PCA) to a reduced feature space of 25 features. The results are based on an Adaboost algorithm and evaluated using the hold-out method. The reported accuracy result is 81.83%.

In a recent work of [32], the authors targeted the problem of classifying attacks using random forest and ANN techniques based on the CICIDS2017 dataset. They applied a package named Boruta for feature selection and returned the top 10 most important features. The feature set is then fed to the classifiers. They reported an average accuracy of 96% using ANN and 96.4% using random forest.

Another dataset that attracted attention in this domain is the UNSW-NB15. In the work of [33], the authors targeted this data set with an approach that uses k-means clustering, CFS feature selection, and four different techniques SVM, RF, J48, and Zero. The proposed approach has been effective in improving the performance of the majority of classifiers. The best-reported accuracy is using J48 with an accuracy of 96.7% and using 10-fold cross-validation.

The authors of [34] targeted also the network intrusion detection problem on the UNSW-NB15. An approach is proposed for the selection of features using a k-means clustering algorithm. The resulting feature set is fed to two classification algorithms: deep neural network (DNN) and random forest. Feature selection has been slightly effective in improving classification performance with 0.03% using DNN and 0.01% using RF. The best-reported accuracy is using DNN of 97.04% using 5-fold cross-validation.

To analyze and summarize the related works, we provide in the next section a taxonomy of IDS and supervised ML algorithms to understand more the reported results and draw up conclusions.

4. Taxonomy of IDS and supervised ML algorithms

In this section, we build a taxonomy of IDS and ML based on the relevant studies mentioned in the preceding section. The taxonomy is based on the following qualities, as given in Table 2: 1) The dataset; 2) Whether or not feature selection is used: Yes or No, 3) Whether or not feature selection is useful: yes or no, 4) The supervised learning algorithms that were employed, 5) The validation method that was used, 6) The highest performing classification algorithm, and 6) The best-reported accuracy and FPR results.

To understand the reported information in Table 2, we will study the datasets separately, and then, we will provide an overall summary:

1. NSL-KDD dataset: on this dataset, whenever feature selection is used, it has been effective in improving classification performance. Moreover, the random forest algorithm is very effective on this dataset and reports high-performance results using different validation methods. ANN seems to be performing well on this data set, however, it was tested on a 20% sample of the data only.
2. KDD'99 dataset: on this dataset, whenever feature selection is used, it has not been effective in improving classification performance. Different validation methods are applied to the data set and no direct conclusion can be drawn about the best performing classification algorithm; however, RF and SVM perform very well on this dataset.
3. CICIDS2017 dataset: no clear inferences about the influence of feature selection can be reached; however, it was beneficial in increasing classification performance in one case. The dataset has a significant imbalance problem, which can be resolved using sampling approaches. QDA is the best classification algorithm.
4. UNSW-NB15 dataset: on this dataset, whenever feature selection is used, it has been effective in improving classification performance. DNN is a common and successful approach to this dataset. It is a big data set and deep learning seems to be the solution.

Overall, we may conclude the following:

1. The problem of IDS utilizing machine learning approaches has received a lot of attention in the literature.
2. Based on a study of four separate data sets, the classification performance is good and promising.
3. Feature selection is critical and, in many circumstances, required for performance improvement.
4. Data imbalance can be a problem, and sampling strategies can help solve the problem.
5. For good performance, a big data set necessitates a deep learning method.

5. Conclusions

Cybercrime is on the rise, thanks to the recent surge in internet content. The use of intrusion detection systems (IDS) is the initial step in detecting and reporting such attacks. The detection of anomalies is reliant on detecting unique attacks, which is a difficult task. This has piqued the interest of scholars all over the world who want to learn more about this subject and, in particular, how to use supervised learning algorithms for intrusion detection to overcome these difficulties.

We go over IDS classification, supervised learning approaches, and cyber security assaults in depth in this study. Then, using four popular datasets: KDD'99, NSL-KDD, CICIDS2017, and UNSW-NB15, we summarized related efforts in this field. Based on these connected works, a taxonomy is also offered. We may deduce from this taxonomy that the field of intrusion detection using supervised machine learning techniques is attracting a lot of interest. In addition, supervised learning algorithms' classification performance is good and promising, according to a study of four data sets: Furthermore, feature selection is critical and, in many circumstances, required for performance improvement. Furthermore, data imbalance can be a concern, and sampling approaches can help resolve the issue. Finally, for good performance, large intrusion detection data sets necessitate a deep learning technique.

Table 2. Taxonomy of IDS

Paper	Dataset	Is feature selection applied?	Is feature selection useful?	Supervised learning algorithms	Validation method	Best performing learning method	Best Reported results
[12]	NSL-KDD	Yes	Yes	J48, Random forest	10-fold	Random forest	Accuracy = 99.7 % FPR = 0.005 %
[24]	NSL-KDD	Yes	Yes	Naive Bayes, Bayes-Net, Logistics, Random forest, J48, Bagging, OneR, PART, ZERO	10-fold	Random forest	Accuracy = 99.9 % FPR = 0.001%
[25]	NSL-KDD	NO	NA	SVM, GMM, Random forest, logistics regression	Hold-out	Random forest	Accuracy = 99 % FPR= NA
[26]	NSL-KDD (20% sample)	YES	Results before feature selection are NA	Artificial neural networks (ANN), SVM	Hold-out	ANN	Accuracy = 94.0% FPR=NA
[27]	KDD’99	YES	Results before feature selection are NA	k-means, random forest, naïve bayes, SVM	Hold-out	Random forest	Accuracy = 99.0% FPR=NA
[28]	KDD’99	YES	NO	SVM	10-fold	SVM	Accuracy=98.7% FPR=NA
[29]	KDD’99	YES	NO	BBNN	10-fold	BBNN	Accuracy=91.0% FPR=NA
[30]	CICIDS2017	YES	YES	Adaboost	Hold-out	Adaboost	Accuracy=81.83% FPR=NA
[31]	CICIDS2017	YES	NA	LDA, QDA, BN, RF	Hold-out	QDA	Accuracy=98.8% FPR=.001%
[32]	CICIDS2017	YES	NA	ANN, RF	Hold-out	RF	Accuracy=96.4
[33]	UNSW-NB15	YES	YES	SVM, J48, RF, Zero	10-fold	J48	Accuracy=96.7% FPR=.13%
[34]	UNSW-NB15	YES	YES	DNN, RF	5-fold	DNN	Accuracy=97.0% FPR=NA

References

[1] Graham, J., Olson, R., & Howard, R. (Eds.). (2011). Cyber security essentials. CRC Press.

[2] Liao, H. J., Lin, C. H. R., Lin, Y. C., & Tung, K. Y. (2013). Intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 36(1), 16-24.

[3] Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. applied sciences, 9(20), 4396

[4] Hamid, Y., Sugumaran, M., & Balasaraswathi, V. R. (2016). Ids using machine learning-current state of the art and future directions. Current Journal of Applied Science and Technology, 1-22.

[5] Masdari, M., & Khezri, H. (2020). A survey and taxonomy of the fuzzy signature-based intrusion detection systems. Applied Soft Computing, 106301.

[6] Milenkoski, A., Vieira, M., Kounev, S., Avritzer, A., & Payne, B. D. (2015). Evaluating computer intrusion detection systems: A survey of common practices. ACM Computing Surveys (CSUR), 48(1), 1-41.

[7] Conrad, E., Misenar, S., & Feldman, J. (2012). CISSP study guide. Newnes

[8] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[9] Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. Acm Sigmod Record, 31(1), 76-77.

[10] https://en.wikipedia.org/wiki/, accessed 1/5/2021

[11] Sahasrabuddhe, A., Naikade, S., Ramaswamy, A., Sadliwala, B., & Futane, P. (2017). Survey on intrusion detection system using data mining techniques. Int Res J Eng Technol, 4(5), 1780-4.

[12] Farnaaz, N., & Jabbar, M. A. (2016). Random forest modeling for network intrusion detection system. Procedia Computer Science, 89, 213-217.

[13] Rust, J. (1997). Using randomization to break the curse of dimensionality. Econometrica: Journal of the Econometric Society, 487-516.

[14] https://www.unb.ca/cic/datasets/index.html, accessed 1-6-2021

[15] www.kaggle.com, accessed 1-6-2021

[16] Uma, M., & Padmavathi, G. (2013). A Survey on Various Cyber Attacks and their Classification. IJ Network Security, 15(5), 390-396.

[17] Li, X., Smith, J. D., Dinh, T. N., & Thai, M. T. (2016, October). Privacy issues in light of reconnaissance attacks with incomplete information. In 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI) (pp. 311-318). IEEE.

[18] Hussain, A., Heidemann, J., & Papadopoulos, C. (2003, August). A framework for classifying denial of service attacks. In Proceedings of the 2003 conference on Applications, technologies, architectures, and protocols for computer communications (pp. 99-110).

[19] Forcht, K. A., Kieschnick, E., Thomas, D. S., & Shorter, J. D. (2007). Identity Theft: The Newest Digital Attack. Issues in Information Systems, 8(2,297-302).

[20] https://en.wikipedia.org/wiki/Cyber_spying#Examples, accessed 20-5-2021

[21] Bhardwaj, A., & Goundar, S. (2020). Keyloggers: silent cyber security weapons. Network Security, 2020(2), 14-19.

[22] Guo, K. H., Yuan, Y., Archer, N. P., & Connelly, C. E. (2011). Understanding nonmalicious security violations in the workplace: A composite behavior model. Journal of management information systems, 28(2), 203-236.

[23] Simmons, C., Ellis, C., Shiva, S., Dasgupta, D., & Wu, Q. (2014, June). AVOIDIT: A cyber attack taxonomy. In 9th Annual Symposium on Information Assurance (ASIA'14) (pp. 2-12).

[24] Malhotra, H., & Sharma, P. (2019). Intrusion Detection using Machine Learning and Feature Selection. International Journal of Computer Network & Information Security, 11(4).

[25] Belavagi, M. C., & Muniyal, B. (2016). Performance evaluation of supervised machine learning algorithms for intrusion detection. Procedia Computer Science, 89, 117-123.

[26] Taher, K. A., Jisan, B. M. Y., & Rahman, M. M. (2019, January). Network intrusion detection using supervised machine learning technique with feature selection. In 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST) (pp. 643-646). IEEE.

[27] El Mourabit, Y., Bouirden, A., Toumanari, A., & Moussaid, N. E. (2015). Intrusion detection techniques in wireless sensor network using data mining algorithms: comparative evaluation based on attacks detection. International Journal of Advanced Computer Science and Applications, 6(9), 164-172.

[28] Li, Y., Xia, J., Zhang, S., Yan, J., Ai, X., & Dai, K. (2012). An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert systems with applications, 39(1), 424-430.

[29] Shah, B., & Trivedi, B. H. (2015, February). Reducing features of KDD CUP 1999 dataset for anomaly detection using back propagation neural network. In 2015 Fifth International Conference on Advanced Computing & Communication Technologies (pp. 247-251). IEEE.

[30] Yulianto, A., Sukarno, P., & Suwastika, N. A. (2019, March). Improving Adaboost-based intrusion detection system (IDS) performance on CIC IDS 2017 dataset. In Journal of Physics: Conference Series (Vol. 1192, No. 1, p. 012018). IOP Publishing.

[31] Abdulhammed, R., Faezipour, M., Musafer, H., & Abuzneid, A. (2019, June). Efficient network intrusion detection using pca-based dimensionality reduction of features. In 2019 International Symposium on Networks, Computers and Communications (ISNCC) (pp. 1-6). IEEE.

[32] Pelletier, Z., & Abualkibash, M. (2020). Evaluating the CIC IDS-2017 Dataset Using Machine Learning Methods and Creating Multiple Predictive Models in the Statistical Computing Language R. Science, 5(2), 187-191.

[33] Hammad, M., El-medany, W., & Ismail, Y. (2020, December). Intrusion Detection System using Feature Selection With Clustering and Classification Machine Learning Algorithms on the UNSW-NB15 dataset. In 2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT) (pp. 1-6). IEEE.

[34] Faker, O., & Dogdu, E. (2019, April). Intrusion detection using big data and deep learning techniques. In Proceedings of the 2019 ACM Southeast Conference (pp. 86-93).